Contribution à l'étude de la régulation transcriptionnelle lors du cylce érythrocytaire de Plasmodium falciparum par l'analyse bioinformatique des acteurs de cette régulation

Charlotte Boschet

UFR de Biologie et Sciences de la Nature
Tour 54/53 – 4ème étage
2 place Jussieu
75005 PARIS

INSERM U511
Immunobiologie cellulaire et moléculaire
des infections parasitaires
91 boulevard de l'Hôpital
75013 PARIS

ECOLE DOCTORALE : BIOCHIMIE ET BIOLOGIE MOLECULAIRE

THESE DE DOCTORAT

pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE DENIS DIDEROT - PARIS 7
Spécialité : ANALYSE DE GENOMES ET MODELISATION MOLECULAIRE

présentée publiquement le 26 juin 2006 par

Charlotte BOSCHET

Contribution à l'étude de la régulation transcriptionnelle lors du cycle érythrocytaire de *Plasmodium falciparum* par l'analyse bioinformatique des acteurs de cette régulation

JURY

Pr Catherine ETCHEBEST, Université Paris 7	Président
Dr Isabelle CALLEBAUT, CNRS	Rapporteur
Pr Gilbert DELEAGE, Université Lyon 1	Rapporteur
Pr Caroline LE VAN KIM, Université Paris 7	Examinateur
Pr Anne-Claude CAMPROUX, Université Paris 7	Examinateur
Dr Catherine VAQUERO, INSERM	Directeur de thèse

A Serge Hazout,
qui devait faire partie de mon jury de thèse
et que j'aurais aimé remercier pour tout ce qu'il a fait pour moi.

Je tiens en tout premier lieu à remercier Isabelle Callebaut, Gilbert Deléage, Anne-Claude Camproux et Caroline Le Van Kim pour avoir accepté cette tâche laborieuse qu'est l'évaluation d'un travail de thèse, ainsi que Catherine Etchebest pour m'avoir si bien conseillé pour la composition de mon jury.

Merci à Dominique Mazier pour m'avoir accueilli au sein de l'unité INSERM U511 mais aussi pour m'avoir permis de développer des talents autres que scientifiques.

Je remercie Catherine Vaquero pour m'avoir donné l'opportunité de faire cette thèse et pour m'avoir laissé une grande liberté dans mon travail.

Je tiens à adresser tous mes remerciements

à Sylvie qui a très vite quitté le statut de collègue pour devenir une amie très sincère. Sylvie, merci pour ton soutien et tes paroles réconfortantes, pour ton amitié et ta confiance, pour tous les bons moments passés et à venir.

à Delphine et Romain pour leur aide inestimable lors de la rédaction de ce manuscrit et leur réconfort. Delphine, j'ai beaucoup aimé enseigner avec toi et j'espère très sincèrement que cela se reproduira dans l'avenir. Romain, merci pour avoir toujours répondu si rapidement à tous les SOS que j'ai pu lancer et merci pour ton amitié sans faille.

à Maryse pour sa bonne humeur, son efficacité, son amitié et toutes les batailles de frites que l'on a pu faire à la piscine et toute la sueur que nous avons évacuée lors des cours de gym.

à Philippe, loin des yeux mais pas loin du cœur.

à tous les membres de l'U511 avec qui j'ai passé de très bons moments sans oublier ceux qui sont partis vers d'autres horizons : Dorothée, Quentin, Nicolas, Stéphane et Ali.

à tous les membres de l'EBGM qui m'ont toujours très bien accueillie et même réconfortée quelques fois. Un merci particulier à Anne qui m'a permis de découvrir une passion le jour où elle m'a demandé si je voulais bien enseigner.

à mes amies les grues : Gud, Ludi et Marie, toujours unies malgré l'éloignement, mes amis les poulpes : Anne, France, Flore, Tiénou et Alexis, et mon californien préféré Alex,

sans oublier Ninou qui a décoré mon bureau de ses dessins et coloriages dédicacés.

Et enfin, mille mercis à mes parents pour m'avoir permis de faire de longues études dans les meilleures conditions possibles. Je n'en serais pas là aujourd'hui s'ils ne m'avaient pas constamment soutenue.

TABLE DES MATIERES

INTRODUCTION

MATERIELS ET METHODES

RESULTATS

ANNEXES

ARTICLES ORIGINAUX

LISTE DES FIGURES ET TABLEAUX

Introduction : Le paludisme et Plasmodium falciparum

Introduction : La transcription et la régulation transcriptionnelle chez les eucaryotes

Matériels et Méthodes

Résultats : Les éléments de régulation

Résultats : Les facteurs de la régulation transcriptionnelle

Annexes

LISTE DES ABREVIATIONS

βME	β-mercaptoéthanol
ADN	acide désoxyribonucléique
ARN	acide ribonucléique
ARNm	ARN messager
ARNt	ARN de transfert
ARNr	ARN ribosomique
bHLH	basic helix-loop-helix
BRE	TFIIB-recognition element
bZIP	basic leucine zipper
C/EBP	CCAAT/enhancer binding protein
CBF	CCAAT-box-binding factor (aussi appelé NF-Y)
CBP	CREB-binding protein
CREB	cAMP-response element binding protein
CSP	circumsporozoite protein
CTF	CCAAT-binding transcription factor (aussi appelé NF-I)
DDT	dichloro-diphényl-trichloroéthane
GBP130	glycophorin-binding protein
GO	Gene Ontology
HAT	histone acétyltransférase
HDAC	histone désacétylase
HMG	high mobility group
HMM	hidden Markov models
HRPII	histidine-rich protein
HTH	helix-turn-helix
Inr	initiateur
KAHRP	knob-associated histidine-rich protein
kb	kilobases
Mb	mégabases
MSP-1	merozoite surface protein
mtTF1	mitochondrial transcription factor
NF-I	nuclear factor-I (aussi appelé CTF)
NF-Y	nuclear factor-Y (aussi appelé CBF)
NHP	non-histone protein
OMS	organisation mondiale de la santé
pb	paires de bases
PCNA	proliferating cell nuclear antigen
RESA	ring-infected erythrocyte surface antigen
RMSD	root mean square deviation
SREBP	sterol regulatory element-binding protein
SSRP	structure-specific recognition protein
TBP	TATA-binding protein
TAF	TBP-associated factors
UBF	upstream binding factor
UPGMA	unweighted pairwise group method average
UTR	untranslated region pour région transcrite mais non traduite

LISTE DES ANGLICISMES ET EXPRESSIONS FRANCISEES

L'anglais est la langue scientifique par excellence. Je présente mes excuses aux farouches défenseurs de la langue française qui liront ce manuscrit mais j'ai préféré garder certains termes anglais ou franciser certaines tournures anglo-saxonnes plutôt que d'utiliser une traduction française approximative ou lourde.

Ainsi vous trouverez dans le texte les termes suivants :

- gap à la place de trou ou insertion-délétion
- position-spécifique, séquence-spécifique, structure-spécifique pour spécifique d'une position, d'une séquence ou d'une structure
- séquence *cis*-régulatrice ou élément *cis*-régulateur plutôt que séquence ou élément agissant en *cis*
- domaine *trans*-activateur pour domaine activant la transcription

AVANT-PROPOS

Le paludisme est une maladie infectieuse transmise par la piqûre d'un moustique. Elle est causée par un parasite eucaryote unicellulaire du genre *Plasmodium*, qui infecte alternativement deux hôtes : l'homme (ainsi que les souris, les oiseaux ou les singes) et la femelle du moustique *Anopheles*, et se caractérise par des accès de fièvre récurrents. Malgré les efforts considérables faits dans les années 1950 et au début des années 1960 pour éradiquer la maladie, il n'y a jamais eu autant de personnes atteintes de paludisme qu'aujourd'hui, notamment en Afrique ; jusqu'à 500 millions de personnes sont infectées dans le monde et plus de 2,7 millions décèdent de la maladie chaque année. La mortalité la plus élevée se situe en Afrique sub-saharienne où 90% des morts dus au paludisme concernent des enfants de moins de 5 ans [40]. De plus les personnes qui survivent souffrent d'anémie et/ou d'immunodépression qui les laisse vulnérables face à d'autres maladies potentiellement mortelles.

Plasmodium falciparum est l'espèce responsable de la forme la plus mortelle de paludisme. Son développement est extraordinairement complexe et nécessite l'expression de certains gènes pour vivre dans son hôte invertébré comme dans son hôte vertébré, pour survivre à l'intérieur des cellules comme à l'extérieur, pour envahir différents types cellulaires comme pour échapper aux cellules du système immunitaire.

Des études menées sur l'expression de différents gènes de cet organisme ont montré que ces gènes semblent répondre au mécanisme transcriptionnel général des eucaryotes. Or, les stratégies de lutte contre ce parasite seront plus efficaces si elles sont ciblées sur une phase précise de son développement et/ou sur des protéines précises exprimées lors de cette phase. Je me suis donc intéressée aux premières étapes de l'expression des gènes du parasite : **la transcription et la régulation transcriptionnelle** et ce, lors de **la phase érythrocytaire de son développement**.

Dans ce manuscrit, je présenterai, dans un premier chapitre d'introduction, la terrible maladie qu'est le paludisme, en terme de victimes mais aussi en terme de conséquences économiques et sociales, ainsi que *P. falciparum*, l'espèce qui provoque les formes les plus graves du paludisme et sur laquelle j'ai travaillé. Dans un deuxième chapitre, je présenterai la transcription et les acteurs de sa régulation chez les eucaryotes.

Dans mon travail, je me suis intéressée plus particulièrement à deux acteurs de la régulation transcriptionnelle qui sont intimement liés car leur interaction est cruciale pour la survie du parasite. D'un côté, les petites séquences d'ADN qui se situent en amont des gènes, c'est-à-dire sur le promoteur, et que l'on appelle des **éléments de régulation** ; de l'autre, les **facteurs de transcription**, qui se fixent soit sur un élément de régulation spécifique soit sur une structure spécifique. La présence, la fréquence et la position de certains éléments de régulation en amont d'un gène, ainsi que la présence ou l'absence de facteurs de transcription à un moment donné du développement font qu'un gène est exprimé ou non. Or lorsque le parasite se développe chez l'homme, notamment dans ses globules rouges, stade responsable des fièvres chroniques de la maladie, on se rend compte, de part les différentes morphologies qu'il adopte, qu'il suit un programme très défini et que chaque gène doit être exprimé à un moment précis du développement.

Les chapitres suivants seront donc consacrés aux résultats obtenus lors de ma thèse concernant dans un premier temps les éléments de régulation et dans un deuxième temps les facteurs de la régulation transcriptionnelle.

INTRODUCTION

LE PALUDISME
ET
PLASMODIUM FALCIPARUM

I - Historique

Premières mentions

Les observations de la « fièvre des marais » remontent à l'Antiquité. Des manuscrits égyptiens datant de 1600 av. J.-C. décrivent les accès paludéens, et établissent une corrélation entre les flambées épidémiques et la saison des pluies. Grecs comme Romains constatent la liaison entre la maladie et la proximité des marécages.

Évoqué dans les écrits du poète Homère, le paludisme est décrit par le médecin Hippocrate : celui-ci mentionne des fièvres sévissant dans les lieux humides, provoquant des frissons et des températures corporelles très élevées à intervalles réguliers, tous les trois ou quatre jours, avec une rate dilatée et douloureuse. En Inde, les signes cliniques de la maladie sont également décrits très tôt.

Le paludisme a indiscutablement influencé l'Histoire en Europe, notamment de la guerre du Péloponnèse qui opposa Athènes et Sparte entre 431 et 404 av. J.-C. et qui vit la déroute de l'armée athénienne en Sicile, lors du troisième conflit, à cause de la maladie, ou encore la fin de la carrière d'Alexandre le Grand qui fut terrassé, à l'âge de 33 ans, par le neuropaludisme en 323 av. J.-C. L'histoire a failli se répéter, encore une fois en Sicile, lorsque les armées alliées débarquèrent conjointement sur l'île le 10 juillet 1943, pour ce qu'on appelle la campagne d'Italie. Un début d'épidémie eut lieu deux semaines après le débarquement, mais les troupes avaient prévu le danger et le pire fut évité grâce à des vaporisations d'arsenic sur les lieux de reproduction des moustiques. Le seul fait de savoir que les moustiques transmettaient le paludisme a sauvé les troupes alliées ; mais les Athéniens auraient dû tenir compte des écrits d'Hippocrate qui, au V^{ème} siècle av. J.-C., avertissait déjà sur les risques de fièvre dans les endroits humides en été.

Premiers traitements

Dans l'Amérique précolombienne, les Amérindiens traitent les fièvres des marais par des infusions d'écorce d'un arbre appelé *Cinchona*. Dans les années 1640, les Jésuites importent la poudre d'écorce de *Cinchona* en Europe, où elle est connue sous le nom de poudre des Jésuites. En 1820, les pharmaciens français Joseph Pelletier et Jean-Baptiste Caventou extraient et identifient chimiquement son principe actif, baptisé quinine. Dans les

années 1830, le médecin militaire français François Clément Maillot codifie son emploi et sa posologie dans les fièvres intermittentes ou continues. La quinine commence également à être utilisée en traitement préventif.

Découverte de la cause du paludisme

Le paludisme, corrélé depuis l'Antiquité aux zones humides, est jusqu'à la fin du XIX^{ème} siècle attribué au « mauvais air » (*mal aria*, en italien) des marais (le mot paludisme vient d'ailleurs du latin *palus* ou *paludis*, « marais »). Au début des années 1880, le médecin français Alphonse Laveran démontre que la maladie est provoquée par un parasite qu'il met en évidence dans les globules rouges de patients contaminés — découverte qui sera récompensée par le Prix Nobel de médecine et physiologie de 1907. Laveran est également le premier à émettre l'hypothèse d'une transmission de ce parasite par les piqûres de moustiques.

Quelques années plus tard, en 1889, le Britannique sir Ronald Ross confirme cette hypothèse et établit le développement du parasite impliqué dans le paludisme des oiseaux — Ross recevra, pour ses travaux, le Prix Nobel de médecine et physiologie de 1902. Enfin, en 1898, l'Italien Giovanni Battista Grassi démontre que les moustiques impliqués dans la transmission du paludisme chez l'homme sont les femelles du genre *Anopheles* et décrit le développement du parasite à l'intérieur de l'organisme des moustiques.

Après la Seconde Guerre mondiale, le paludisme est éradiqué dans les régions d'Europe tempérée et d'Amérique du Nord où il sévit, grâce à l'épandage massif de dichloro-diphényl-trichloroéthane (DDT) et à de larges campagnes d'assèchement des marais.

Découvertes ultérieures

La quinine reste pendant longtemps le seul traitement disponible contre le paludisme. Mais en 1930, de nouvelles molécules font leur apparition : ce sont les premiers antipaludéens de synthèse, avec la chloroquine (1934). D'autres molécules sont régulièrement découvertes tout au long du XX^{ème} siècle, mais l'histoire du paludisme est alors marquée par l'apparition de souches de parasites résistantes aux médicaments utilisés de façon massive.

En 1972, des scientifiques chinois isolent le principe actif de l'armoise annuelle (Qinghaosu ou *Artemisia annua*), une plante traditionnellement utilisée en Chine pour combattre le paludisme. Son usage comme remède contre les fièvres remonterait à l'an 340 et son emploi contre les symptômes du paludisme aurait commencé vers le milieu du XVI^{ème} siècle. Contrairement aux autres traitements, aucune résistance n'a été observée face à ce composant actif, baptisé artémisine. Cette efficacité a conduit l'Organisation Mondiale de la Santé (OMS) à conclure en 2001 un accord avec le gouvernement chinois pour produire en grande quantité un traitement contre le paludisme à base de ce composé. Mais aujourd'hui, son prix le met hors de portée de la plupart des malades africains.

En 2002, la connaissance du parasite, le plasmodium, et des moustiques vecteurs, les anophèles, franchit une nouvelle étape avec l'annonce, par un consortium international de scientifiques [118, 169], du séquençage complet des génomes de *Plasmodium falciparum* et de *Anopheles gambiae* (la principale espèce d'anophèle vecteur du paludisme en Afrique).

II - Le paludisme et les hommes

II.1 - Symptômes et complications

La gravité du paludisme dépend du type de plasmodium impliqué, de la quantité de parasites dans le sang, et du sujet lui-même (âge, degré d'immunisation, …). Le paludisme est ainsi très sévère chez les enfants entre 3 mois et 4 ans (protégés par les anticorps maternels pendant les premiers mois de leur vie, ils ne commencent à fabriquer leurs propres anticorps qu'aux alentours de 4-5 ans).

Accès palustres

La primo-invasion se caractérise par une phase silencieuse d'incubation de durée variable, entre 8 jours et un mois — cependant, chez les voyageurs ayant suivi un traitement préventif, la poussée de paludisme peut survenir de plusieurs mois à un an après l'infestation (dans le cas de *P. vivax*). Ensuite apparaissent les premiers symptômes, qui consistent en une poussée de fièvre élevée (40 à 41 °C), parfois accompagnée de maux de

tête, de douleurs musculaires, d'un affaiblissement général, voire de vomissements et de diarrhées. En région tropicale (ou au retour d'un voyage), toute apparition de fièvre, même en l'absence d'autres symptômes, doit être au premier abord considérée comme un paludisme.

Après l'accès fébrile de la première invasion, il existe un risque de passage aux accès intermittents. Ce sont des épisodes de crises paludéennes dans lesquelles se succèdent de façon typique une phase de frissons intense, une phase de fièvre puis une phase de sueurs froides (baisse de la température et transpiration abondante). Ces épisodes, appelés accès palustres, correspondent à l'éclatement des globules rouges en raison de la multiplication des parasites ; ils apparaissent selon un rythme régulier dont la périodicité dépend de l'espèce de plasmodium impliquée : tous les deux jours (fièvre tierce bénigne due à *Plasmodium vivax* et *Plasmodium ovale*, fièvre tierce maligne due à *Plasmodium falciparum*) ou tous les trois jours (fièvre quarte due à *Plasmodium malariae*). Il peut toutefois exister des fièvres quotidiennes en cas de double infestation dans laquelle les cycles des parasites sont décalés, ou en cas de fièvre tierce maligne, qui présente des symptômes plus atypiques (la succession frisson-fièvre-sueurs froides y est moins nette).

Évolution en l'absence de complication

Chez les personnes qui vivent en région d'endémie, le risque d'infections successives, accompagnées de fréquentes récidives de la maladie est élevé. Ces patients finissent cependant assez fréquemment par être immunisés contre les souches de plasmodium à laquelle ils sont régulièrement confrontés.

Chez les personnes qui quittent la région d'endémie et en l'absence de traitement, la maladie finit généralement par se résorber de façon spontanée. Dans les cas non compliqués, elle disparaît ainsi en deux à trois mois pour *Plasmodium falciparum*. Chez les autres espèces de plasmodiums, la formation de formes dormantes de parasites dans le foie, alors appelés hypnozoïtes, conduit à la possibilité de rechutes et de persistance du parasite dans l'organisme pendant des périodes beaucoup plus longues : deux à trois ans pour *Plasmodium vivax*, cinq ans environ avec *Plasmodium ovale* et de 10 à 20 ans, voire jusqu'à 30 ans, pour *Plasmodium malariae*.

Complications

Les complications potentielles du paludisme sont liées dans la majorité des cas à une infestation par *Plasmodium falciparum* — les autres espèces de paludisme provoquant des formes bénignes (même si elles peuvent persister plusieurs années) de la maladie.

‣ L'accès pernicieux ou neuropaludisme

Chez les sujets non immunisés ou ne suivant pas de traitement, l'infection à *Plasmodium falciparum* présente un risque de développement d'une forme grave potentiellement mortelle : le neuropaludisme, responsable d'une grande partie de la mortalité infantile liée au paludisme. Il se traduit en particulier par des altérations de la conscience, des délires, des convulsions, pouvant aboutir à un coma et à la mort. Les mécanismes du neuropaludisme ne sont pas encore élucidés ; l'une des hypothèses est le blocage des petits vaisseaux sanguins (capillaires) du cerveau par des amas de globules rouges infestés. Le neuropaludisme constitue une urgence médicale.

‣ Le paludisme viscéral évolutif

Le paludisme viscéral est une complication assez rare qui peut survenir avec *Plasmodium falciparum* et, dans une moindre mesure, *Plasmodium vivax*. Il apparaît à la suite d'infestations successives et massives mal ou non traitées chez des sujets non immunisés, ou ayant perdu leur immunisation (par exemple chez les natifs de zones d'endémies quittant ces régions pendant de longues périodes et y retournant de façon ponctuelle). Le paludisme viscéral évolutif associe notamment pâleur, fatigue intense (asthénie), anémie, splénomégalie (augmentation du volume de la rate), fièvres irrégulières. En l'absence de traitement, il existe un risque permanent de neuropaludisme (en cas d'infestation par *Plasmodium falciparum*).

‣ Le paludisme chez la femme enceinte

L'infection par un plasmodium chez la femme enceinte a des conséquences très sévères, en particulier si l'infection a lieu pendant le premier ou le troisième trimestre de la grossesse : elle peut se traduire par un avortement spontané ou la mort néonatale. Dans les cas moins sévères, elle s'accompagne de risques élevés de prématurité, ou de mise au monde d'un enfant de faible poids.

II.2 - Prévention

La protection totale contre le paludisme est impossible ; la réduction du risque passe par l'évitement, dans la mesure du possible, des piqûres de moustiques : usage de moustiquaires et de répulsifs anti-moustiques, port de pantalons et vêtements couvrants et de chaussures fermées pendant les périodes d'activité des moustiques.

La prévention médicamenteuse consiste en la prise d'un traitement antipaludéen pendant les séjours en zones endémiques. Elle vise à éviter le développement de la maladie en cas d'infection, mais n'est pas capable d'empêcher l'infection en cas de piqûre par un moustique contaminé. Par ailleurs, le traitement préventif n'offre pas une protection totale : il ne dispense donc pas de la protection contre les piqûres de moustiques.

II.3 - Traitements

Les traitements contre le paludisme portent le nom générique d'antipaludéens. Les phénomènes de résistance des plasmodiums aux médicaments antipaludéens sont en constante progression, en particulier chez *Plasmodium falciparum*. Le choix d'une molécule antipaludéenne doit donc non seulement tenir compte, comme pour tout médicament, de la tolérance du patient, mais également des phénomènes de résistance des souches parasitaires.

Types d'antipaludéens

Les antipaludéens peuvent être classés selon la localisation de leur action dans le développement des plasmodiums, qui passent par les stades successifs suivants : sporozoïte (forme injectée par le moustique), mérozoïte (forme libérée par le foie), schizonte (forme de multiplication dans les globules rouges) et gamétocyte (future cellule reproductrice).

On distingue sur cette base deux grands types d'antipaludéens : les schizontocides, qui agissent sur les schizontes, et les gamétocytocides, actifs sur les gamétocytes. Les premiers permettent de lutter contre les symptômes du paludisme, les seconds de contrer la transmission du parasite. Les molécules qui s'attaquent aux mérozoïtes sont pour l'instant très peu utilisées en raison de leur forte toxicité pour le foie.

Principales molécules

Les principales molécules antipaludéennes agissent en bloquant certaines réactions métaboliques du parasite et peuvent donc être classées selon leur mode d'action. Les médicaments antipaludéens contiennent une molécule ou deux, en association. En France, ils ne sont délivrés que sur prescription médicale.

‣ Les schizonticides électifs

Ce groupe comprend les dérivés des deux seules substances que l'on trouve à l'état naturel, la quinine et l'artémisine :

- la quinine et ses dérivés (chloroquine, méfloquine, halofantrine) se fixent sur l'hème libéré par la digestion de l'hémoglobine du globule rouge parasité, source d'acides aminés. Ils empêchent ainsi sa transformation en hémozoïne, un pigment insoluble et inoffensif. L'accumulation d'hème non transformé est alors toxique pour la cellule et entraîne la lyse du parasite ;
- l'artémisine et ses dérivés synthétiques agissent très rapidement par rapport aux autres molécules. Au départ, on pensait que cette substance libérait des radicaux libres qui à leur tour attaquaient et brisaient la membrane cellulaire du parasite. Mais, selon les travaux publiés en 2003 [100, 323], l'artémisine agirait en fait en bloquant l'action d'une enzyme (PfATP6) essentielle pour pomper le calcium de et vers les cellules du parasite. Toutes les cellules complexes ont besoin de ces pompes pour leur moteur moléculaire.

‣ Les inhibiteurs des acides nucléiques ou antimétabolites

Ils bloquent la division du noyau du parasite. Ce groupe comprend les antifolates, les naphtoquinones et les antibiotiques :

- les antifolates répartis en deux familles : les antifoliques (sulfadoxine) et les antifoliniques (proguanil, pyriméthamine). Ils agissent au niveau de la voie de synthèse des folates, qui sont essentiels à la biosynthèse des acides nucléiques et donc indispensables à la synthèse d'ADN. Les antifoliques inhibent la dihydroptéroate synthétase (DHPS) qui produit l'acide folique, les antifoliniques inhibent la dihydrofolate réductase (DHFR) qui produit l'acide folinique ;

- les naphtoquinones : l'atovaquone est un inhibiteur puissant des fonctions mitochondriales. Elle bloque la chaîne de transfert d'électrons au niveau de son enzyme-clé, la dihydroorotate déshydrogénase (DHOdase). Elle a peu d'impact thérapeutique lorsqu'elle est utilisée seule. En combinaison avec un antimétabolite comme le proguanil, on observe une intéressante synergie d'action grâce à une inhibition séquentielle de la synthèse des pyrimidines. Une originalité de l'association atovaquone-proguanil est son action sur les stades hépatocytaires de *P. falciparum* ;
- les antibiotiques : les tétracyclines (doxycycline), les macrolides (érythromycine, azythromycine, clindamycine) peuvent inhiber la synthèse protéique par inhibition de certaines fonctions de l'apicoplaste, organelle vestige d'une symbiose avec un procaryote.

Phénomènes de résistance

L'utilisation massive de molécules antipaludéennes a entraîné — et continue d'entraîner — l'apparition de souches de plasmodiums résistantes à ces traitements. Cette résistance est en hausse constante, en particulier en Afrique sub-saharienne ; ainsi l'emploi de la chloroquine, massif il y a une vingtaine d'année, est-il désormais limité en raison de l'apparition de plasmodiums résistants dans plusieurs régions d'Afrique, d'Asie du Sud-Est et d'Amérique du Sud.

Des souches résistantes sont également apparues contre l'halofantrine et la méfloquine. Les résistances à la quinine existent, mais restent de niveau peu élevé. Enfin, il existe également des souches de plasmodiums multirésistantes, c'est-à-dire insensibles à plusieurs des molécules antipaludéennes disponibles. Pour éviter l'installation de la résistance, on utilise généralement des combinaisons de molécules qui n'ont pas le même mode d'action.

L'OMS tient à jour la liste des pays et des régions dans lesquelles il existe des souches de plasmodiums résistantes, et à quelles molécules, ce qui permet de proposer aux patients des traitements adaptés. L'OMS définit ainsi trois grandes zones géographiques, A, B et C (Figure 1) :

- zone A : pays où le risque est faible et saisonnier, et où *Plasmodium falciparum* est absent ou bien sensible aux antipaludéens ;

- zone B : pays où le risque est faible, où *Plasmodium vivax* est sensible à la chloroquine et où *Plasmodium falciparum* est relativement sensible à la combinaison chloroquine/proguanil ;
- zone C : pays où *Plasmodium falciparum* est résistant à la chloroquine ou à d'autres molécules.
-

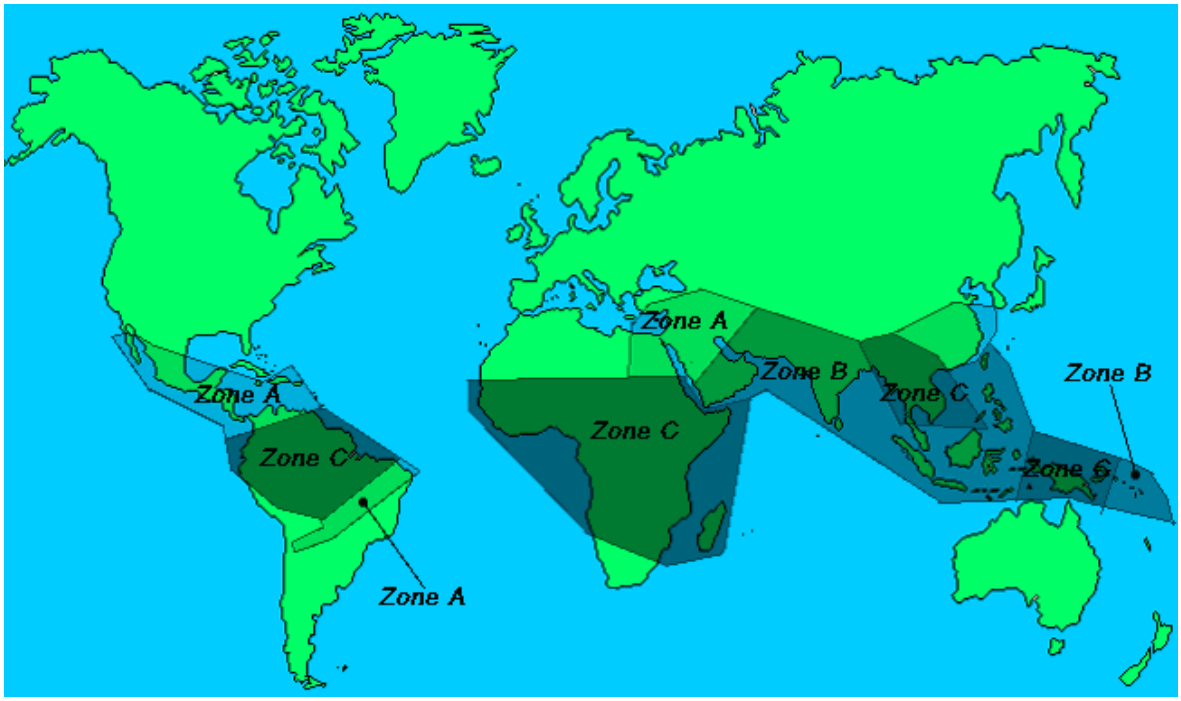


Figure 1. Zones à risque définies par l'OMS.

En France, le classement émis par le Ministère de la Santé définit les groupes 0, I, II et III ; il ne recoupe pas complètement la classification de l'OMS :

- groupe 0 : pays où le paludisme est absent ;
- groupe I : pays où le paludisme existe, mais où *Plasmodium falciparum* n'existe pas ;
- groupe II : pays où il existe des phénomènes de résistance modérée ;
- groupe III : pays où l'on trouve des souches de *Plasmodium falciparum* résistantes ou multirésistantes.

Traitements curatifs et préventifs

Les molécules utilisées en prévention et en traitement sont les mêmes, seules les posologies diffèrent : elles sont plus faibles en usage préventif.

‣ Traitements préventifs

Les traitements préventifs ont pour but d'empêcher le développement de la maladie en cas d'infection ; ils n'empêchent en aucun cas l'infection en cas de piqûre par un moustique porteur. Ils sont prescrits lors de voyages dans les régions où le paludisme est endémique — une consultation médicale est indispensable. Cette chimioprophylaxie débute généralement la veille du départ ou une semaine avant selon le médicament utilisé, se poursuit pendant la durée du séjour et se termine de quatre à six semaines après le retour. La protection conférée par les antipaludéens pris à titre préventif n'est pas absolue. Il est donc indispensable, parallèlement, de se protéger des piqûres de moustiques.

‣ Traitements curatifs

La quinine reste la molécule la plus utilisée dans le traitement du paludisme. Elle constitue également le médicament d'urgence indispensable (elle est alors administrée par voie intraveineuse). D'autres molécules peuvent également être prescrites en fonction de la souche de plasmodium impliquée et du patient lui-même (certains traitements sont en effet contre-indiqués en cas de troubles digestifs ou de grossesse).

Le traitement par antipaludéens est assez bien toléré et les effets indésirables (céphalées, nausées, vertiges) sont rares en utilisation préventive, et bénins en utilisation curative.

II.4 - Le paludisme dans le monde

Un problème majeur de santé publique

Le paludisme est, avec le sida et la tuberculose, l'une des trois principales causes de mortalité d'origine infectieuse. L'OMS estime qu'il touche entre 300 et 500 millions de personnes dans le monde, dont 90% en Afrique sub-saharienne — les autres cas survenant en Asie du Sud-Est et en Amérique du Sud (Figure 2). Il cause entre 1,5 et 2,7 millions de décès par an. L'Afrique est particulièrement touchée : la maladie y tue un enfant toutes les 30 secondes (plus de 1 million de décès infantiles par an).

Dans les pays exempts de paludisme, il existe toutefois un paludisme dit d'importation, qui touche les personnes rentrant de voyages en régions d'endémie (pour environ 95% des

cas en Afrique sub-saharienne). En France métropolitaine, le paludisme d'importation concerne environ 7 000 personnes par an [76].

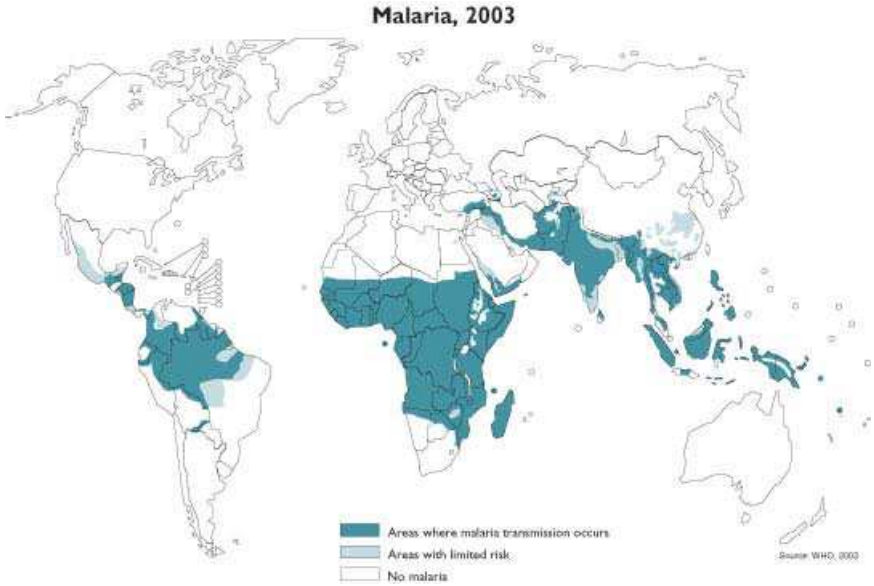


Figure 2. Pays touchés par le paludisme (Source : OMS, 2003).

Lutte contre le paludisme

Bien que le XX^{ème} siècle ait vu l'émergence, dans divers pays, de plusieurs programmes d'élimination du parasite -programmes ayant eu des résultats positifs-, cette maladie devient un fardeau de plus en plus lourd pour le monde. Ceci est attribué à plusieurs causes comme les mouvements des populations dans les régions endémiques, les changements dans les pratiques agricoles (construction de barrages et de plans d'irrigation), la déforestation, les faibles moyens donnés aux systèmes de santé publique dans certains pays en voie de développement, ou encore les changements climatiques comme le phénomène El Niño [1] ou le réchauffement de la planète. De plus, la résistance aux médicaments et insecticides utilisés pour neutraliser la maladie a évolué avec le nombre grandissant de cas cliniques. Avec une population augmentant rapidement dans les régions où la transmission de la maladie est la

[1] L'expression El Niño (signifiant "l'Enfant Jésus" en espagnol) était utilisé à l'origine par les pêcheurs le long des côtes de l'Équateur et du Pérou et s'appliquait à un courant océanique chaud qui apparaît habituellement au moment de Noël pour ne disparaître que quelques mois plus tard. Les poissons sont alors moins abondants pendant ces intervalles chauds, et les pêcheurs souvent en profitent pour réparer leur équipement de pêche et rester avec leurs familles. Certaines années, cependant, l'eau est particulièrement chaude, et l'arrêt de la saison de pêche s'éternise jusqu'à mai ou quelquefois juin. Avec le temps, l'utilisation de l'expression "El Niño" a été réservée à ces intervalles exceptionnellement chauds et marqués, qui non seulement perturbent les vies de ces pêcheurs sud-américains, mais également, apportent des pluies intenses : un climat chaud et humide idéal pour la transmission du paludisme.

plus forte, on estime que le nombre de cas de paludisme doublera dans les 20 prochaines années si aucune stratégie efficace n'est mise en place [40].

La lutte contre le paludisme comprend trois volets complémentaires : le traitement des malades, la prévention et la protection contre les anophèles, vecteurs de la maladie. En 1955, l'OMS a mis en place un programme global d'éradication du paludisme, qui manque cependant des moyens financiers nécessaires et qui peut se heurter à certains gouvernements.

Outre l'épandage d'insecticides, qui peuvent être toxiques pour les écosystèmes (c'est ainsi que le DDT est aujourd'hui interdit dans de nombreux pays) et provoquer l'apparition d'insectes résistants, la lutte contre les moustiques passe par l'assèchement des marais ou leur transformation en eaux courantes et la destruction des points d'eau stagnante (sites privilégiés de reproduction des anophèles), en particulier aux alentours des habitations. Ces mesures sont toutefois difficiles à appliquer dans les pays où sévit la sécheresse ou dans les régions où les installations sanitaires sont déficientes.

En 1998, l'OMS a lancé un nouveau programme, appelé Roll Back Malaria, en partenariat avec la Banque Mondiale et les Nations Unies, qui vise à faire reculer notablement le paludisme d'ici 2010.

Le paludisme et les autres maladies

Un des résultats surprenants de l'utilisation à grande échelle de moustiquaires imprégnées d'insecticides est que la réduction de la mortalité est plus grande que la réduction de la mortalité uniquement due au paludisme. Ceci implique que le paludisme est intimement lié à d'autres maladies comme facteur direct, ou alors que le paludisme rend la population plus susceptible aux autres infections [349].

Les effets indirects de la maladie commencent bien avant la naissance. Les femmes enceintes ont un risque très élevé d'être infectées par le paludisme, du fait de leur immunité diminuée, et de grossesses impaludées peuvent résulter des fausses couches, des décès infantiles, une diminution du poids à la naissance ou encore des infections congénitales. Des infections palustres sévères ou chroniques peuvent altérer le système immunitaire et la réponse aux vaccins et ainsi augmenter la vulnérabilité aux autres infections.

En outre, le paludisme chronique est un important facteur de l'anémie [164, 341], qui a des effets physiques directs et qui diminue la productivité du travailleur [21, 338]. Le paludisme est aussi associé à la splénomégalie, à des affections rénales chroniques et au syndrome néphrétique, ainsi qu'au lymphome de Burkitt. Le paludisme devient de plus en plus un facteur important dans la transmission du virus de l'immunodéficience humaine (VIH) car les enfants avec une sévère pathologie requièrent de nombreuses transfusions sanguines et une bonne partie des réserves de sang en Afrique sub-saharienne est infectée par le VIH. Donc comme le paludisme est un facteur important d'autres maladies, toute évaluation du poids économique du paludisme doit inclure les coûts associés à toutes ces autres maladies.

III - *Plasmodium falciparum*

Les parasites eucaryotes unicellulaires du genre *Plasmodium* appartiennent au phylum des apicomplexes. Ce phylum forme un groupe hétérogène et les différents ordres qui le composent ont surtout en commun leur mode de vie endoparasitaire et une combinaison d'organites formant le complexe apical. Les plasmodiums sont des parasites des hématies, c'est pourquoi on les nomme parfois Hématozoaires. Ils appartiennent plus particulièrement à l'ordre des Hémosporidies et à la famille des Plasmodiidés (Tableau 1).

Tableau 1. Classification des Apicomplexa et de certains organismes eucaryotes unicellulaires.

Règne	Phylum	Classe	Ordre	Famille	Genre
Protiste	Apicomplexa	Haemosporidea	Haemosporida	Plasmodiidae	*Plasmodium* (*)
		Coccidea	Eimeriida	Eimeriidae	*Eimeria*
				Cryptosporidiidae	*Cryptosporidium*
				Sarcocystiidae	*Sarcocystis*
					Toxoplasma
		Piroplasmidea	Piroplasmida	Babesiidae	*Babesia*
				Theileriidae	*Theileria*
	Amoebozoa	Mycetozoa	Dictyosteliida	Dictyosteliidae	*Dictyostelium* (*)
		Oligohymenophorea	Entamoebida	Entamoebidae	*Entamoeba* (*)
	Microspora	Microsporididea	Pleistophoridida	Pleistophoridae	*Encephalitozoon*
Champignon	Ascomycota	Saccharomycetes	Saccharomycetales	Saccharomycetaceae	*Saccharomyces*
		Schizosaccharomycetes	Schizosaccharomycetales	Schizosaccharomycetaceae	*Schizosaccharomyces*

Les astérisques (*) indiquent les genres dans lesquels certaines espèces ont un génome très riche en A+T qui a été entièrement séquencé : *Plasmodium falciparum* (80,6% en moyenne), *Dictyostelium discoideum* (77,2%), *Entamoeba histolytica* (75,3%).

Il existe de nombreuses espèces de *Plasmodium*, dont quatre provoquent des formes plus ou moins sévères du paludisme chez l'homme (Tableau 2) :

- *Plasmodium vivax*, le plus répandu, est présent dans le monde entier (c'est l'espèce qui sévit dans le bassin méditerranéen). Il est responsable de formes bénignes du paludisme (fièvres tierces bénignes) et n'entraîne que rarement des complications ;
- *Plasmodium falciparum* est également cosmopolite ; c'est l'espèce qui provoque les formes les plus graves du paludisme (notamment l'accès pernicieux ou neuropaludisme) ;
- *Plasmodium malariae* est beaucoup plus rare. Cosmopolite, mais se rencontrant principalement en Afrique tropicale, il est responsable de la fièvre quarte ;
- *Plasmodium ovale* est l'espèce la plus rare ; on le trouve en Afrique centrale et occidentale. Il provoque des formes bénignes du paludisme (fièvres tierces bénignes).

Tableau 2. Comparatif des quatre espèces de parasites *Plasmodium* infectant l'homme.

Espèces		*Plasmodium vivax*	*Plasmodium falciparum*	*Plasmodium malariae*	*Plasmodium ovale*
Cycle pré-érythrocytaire (jours)	(1)	8	5,5 - 6	13	9
Période prépatente (jours)	(2)	11 - 13	9 - 10	15 - 16	10 - 14
Période d'incubation (jours)	(3)	12 - 17 ou jusqu'à 6 - 12 mois	9 - 14	18 - 40 ou plus	16 - 18 ou plus
Second cycle exo-érythrocytaire	(4)	oui	non	dans certaines souches	oui
Cycle érythrocytaire (heures)		48	48	72	49 - 50
Parasitémie moyenne (par mm3)		20 000	20 000-500 000	6 000	9 000
Parasitémie maximale		50 000	2 000 000	20 000	30 000
Sévérité de la première attaque	(5)	moyenne à sévère	sévère	moyenne	moyenne
Rechutes possibles		++	-	+++	++
Période de réapparition de la maladie		longue	courte	très longue	longue
Durée de l'infection (années)		1,5 - 3	1 - 2	3 - 50	1,5 - 3

(1) période qui débute au moment où le sporozoïte est injecté pour la première fois dans la circulation sanguine par le moustique jusqu'à ce que les mérozoïtes soient relâchés par le schizonte hépatique et infectent un globule rouge, (2) intervalle entre l'infection et le moment où les parasites sont détectables dans le sang périphérique, (3) intervalle entre l'infection et l'apparition des symptômes, (4) les mérozoïtes produits par un schizonte hépatique peuvent réinfecter des cellules hépatiques, (5) chez les personnes immunodéficitaires pour *Plasmodium falciparum*.
Données obtenues sur http://www.brown.edu/Courses/Bio_160/Projects1999/malaria/ldpg.html.

III.1 - Le développement de *Plasmodium falciparum*

Plasmodium falciparum nécessite deux hôtes successifs pour se développer : l'homme qu'il infecte et un moustique qui lui sert de vecteur sain. A l'intérieur de ses deux hôtes consécutifs, le parasite suit un développement complexe qui comporte deux étapes essentielles : une phase asexuée chez le moustique puis chez l'homme, et une phase sexuée qui commence dans les érythrocytes humains et se termine chez le moustique (Figure 3).

Au cours de son repas sanguin, le moustique infecté injecte, dans la circulation sanguine de l'homme, le parasite sous forme de sporozoïtes, où il reste de quelques minutes à une demi-heure avant de disparaître. Beaucoup d'entre eux sont phagocytés par des macrophages, mais quelques-uns envahissent les hépatocytes où ils subiront la première phase de leur développement : la **schizogonie pré-érythrocytaire** ou hépatocytaire. Les sporozoïtes deviennent alors des trophozoïtes hépatiques qui se multiplient intensément dans une même vacuole parasitophore jusqu'à devenir un schizonte hépatique multinucléé. Une semaine après la piqûre de l'anophèle, les schizontes éclatent et libèrent dans la circulation sanguine des milliers de mérozoïtes. Cette phase hépatique est asymptomatique, on parle alors d'incubation. Intervient ensuite la deuxième phase du développement parasitaire, quand les mérozoïtes pénètrent dans des globules rouges : la **schizogonie érythrocytaire**. Cette étape est, comme l'étape précédente, une étape de prolifération intense. Les mérozoïtes sont alors au stade anneau ; en se multipliant, ils deviennent des trophozoïtes puis des schizontes érythrocytaires. Après deux ou trois jours, le globule rouge éclate libérant ainsi 8 ou 16 nouveaux mérozoïtes qui pourront alors infecter de nouveaux globules rouges. Ce sont ces cycles de multiplication dans le sang, et plus particulièrement l'explosion des globules rouges, qui provoquent les fièvres périodiques caractéristiques du paludisme ; plus le nombre de parasites augmente, plus la personne infectée est malade.

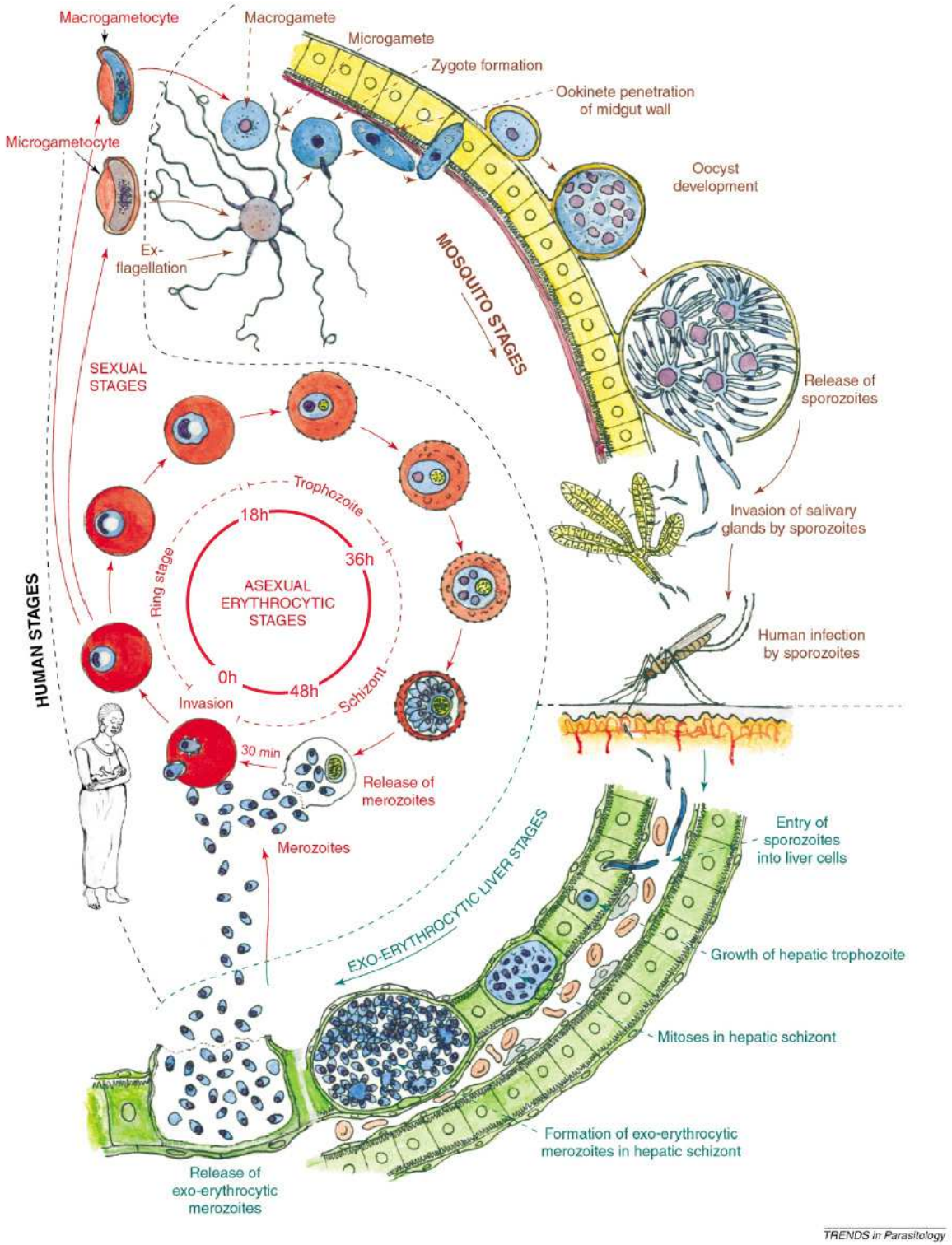


Figure 3. Le développement de *Plasmodium falciparum* dans ses deux hôtes.
Les principaux stades du développement du parasite dans le foie et les globules rouges (stades asexué et sexué) chez l'homme ainsi dans le tube digestif et les glandes salivaires chez le moustique sont représentés ici. Pour plus de détails, voir le texte. Image tirée d'un article de L. Bannister & G. Mitchell [19]

Néanmoins, il existe une alternative pour certains mérozoïtes : l'entrée dans la **gamétocytogénèse**. Les parasites se transforment en gamétocytes mâles (macrogamètes) ou femelles (microgamètes) qui ne continueront leur différenciation que s'ils sont « absorbés » par un moustique lors de son repas sanguin. Après fécondation dans les intestins du moustique, le zygote ainsi formé devient en quelques heures un ookinète mobile. Il entamera ensuite la dernière phase de son développement : la **sporogonie**. Dans la membrane stomacale, le parasite grandit rapidement, formant un oocyste sphérique. Au bout d'une semaine, selon la température ambiante, l'oocyste commence un processus de divisions internes, avec formation de sporozoïtes en forme de petits vers. L'oocyste explose 14 jours après le début de ce processus de division. Les sporozoïtes ainsi libérés gagnent ensuite les glandes salivaires du moustique et se déverseront dans le sang circulant de la prochaine victime du moustique. Un cycle nouveau pourra de ce fait commencer.

Pour résumer, *Plasmodium* subit plusieurs phases de prolifération intense aussi bien chez le moustique lors de la sporogonie que chez l'homme, pendant la schizogonie exo-érythrocytaire et la schizogonie érythrocytaire, cette dernière phase étant la phase responsable de la maladie. Il passe aussi par une phase de différenciation cellulaire, associée à un arrêt de la prolifération : la gamétocytogénèse, qui, elle, est responsable de la dissémination de la maladie.

Dans mon travail, je me suis essentiellement intéressée à la phase érythrocytaire parce que c'est la phase responsable des symptômes : les fièvres correspondent à l'explosion des érythrocytes et la libération des mérozoïtes dans la circulation sanguine.

III.2 - Le génome de *Plasmodium falciparum*

En 1996, un consortium international [118, 169] réunissant des scientifiques de trois grands groupes : le Sanger Centre à Cambridge, The Institute of Genomic Research (TIGR) et l'Université de Stanford aux Etats-Unis, a entrepris de séquencer le génome nucléaire complet du clone 3D7 de *Plasmodium falciparum*. Le but de ce projet était de connaître de manière précise la biologie du parasite de façon à trouver des failles dans l'armure du parasite afin de développer de nouveaux médicaments et vaccins efficaces. Lorsque j'ai

commencé ce travail, le séquençage n'était pas terminé et l'on avait accès à des morceaux de séquences sous forme de shotguns ou de contigs.

Au moment de la publication du génome en octobre 2002 [132], les chromosomes 1 à 5, 9 et 12 étaient entièrement séquencés, alors que les chromosomes 6 à 8, 10, 11, 13 et 14 contenaient encore de 3 à 37 « trous », la plupart de moins de 2,5 kilobases (kb).

Les caractéristiques du génome nucléaire

Le génome nucléaire du clone 3D7 de *Plasmodium falciparum* (Tableau 3) est composé de 22,85 mégabases (Mb) distribuées en 14 chromosomes ayant une longueur comprise entre 0,643 et 3,29 Mb. Ainsi, le génome de *P. falciparum* est deux fois plus long que les génomes des levures *Schizosaccharomyces pombe* et *Saccharomyces cerevisiae* (Figure 4). Sa composition générale en (A+T) est de 80,6% et peut approcher les 90% dans les introns et les régions intergéniques.

Les phases ouvertes de lecture ont été prédites par divers programmes informatiques et contrôlées manuellement. On a identifié sur ce génome environ 5 300 séquences codantes, ce qui se situe entre ce qui a été prédit chez *S. pombe* et *S. cerevisiae* (Tableau 3 & Figure 4).

La densité de gènes est ainsi d'environ 1 gène toutes les 4 338 pb ce qui est légèrement plus que ce qui avait été calculé lors de la publication des chromosomes 2 et 3 (respectivement 1 gène pour 4 500 pb et 1 gène pour 4 800 pb) [38, 133].

Des introns ont été prédits dans 54% des gènes. Et en excluant les introns, les gènes ont une longueur moyenne de 2,3 kb ce qui est légèrement plus long que dans les autres génomes d'eucaryotes unicellulaires présentés ici, dont la longueur moyenne des gènes va de ~1,4 à ~1,8 kb. *Plasmodium falciparum* montre une plus grande proportion de gènes (15,5%) dont la taille est supérieure à 4 kb par comparaison à *S. cerevisiae* et *S. pombe* (respectivement 3,0% et 3,6%). La raison de cette augmentation reste incertaine. Beaucoup de ces grands gènes codent des protéines encore non caractérisées qui semblent être cytosoliques car elles ne présentent pas de peptide signal reconnaissable [132]. Cependant, dans les protéines déjà caractérisées, la présence d'insertions entre les domaines caractéristiques a aussi été observée [306], notamment dans les ARN polymérases I, II et III [123, 231, 232] qui sont des protéines nucléaires.

Tableau 3. Résumé des caractéristiques du génome de *P. falciparum* et comparaison avec d'autres génomes d'organismes eucaryotes unicellulaires.

Caractéristiques	Valeurs					
	P. falciparum	*C. parvum*	*D. discoideum*	*E. cuniculi*	*S. cerevisiae*	*Sc. pombe*
Nombre de chromosomes	14	8	6	11	16	3
Longueur (pb)	22 853 764	9 087 724	34 042 810 ‡	2 507 519	12 495 682	12 462 637
Nombre de gènes	5 268	3 807	12 500 ‡	1 997	5 770	4 929
Longueur moyenne des gènes (pb) †	2 283	1 795	1 756	-	1 424	1 426
Longueur moyenne des régions intergéniques (pb)	1 694	-	-	129	515	952
Densité de gènes (pb par gène)	4 338	2 382	2 500	1 256	2 088	2 528
Pourcentages de séquences codantes	52,6%	75,3%	-	90,0%	70,5%	57,5%
%GC						
Total	19,4	30,0	22,4	47,0	38,3	36,0
Exons	23,7	-	27,0	47,6	28,0	39,6
Introns	13,5	-	12,0	-	49,0	-
Régions intergéniques	13,6	23,9	15,0	45,0	35,1	32,4
ARN						
Nombre de gènes d'ARNt	45	45	390	44	299	174
Nombre de gènes d'ARNr 5S	3	6	1 #	3	100-200	30
Nombre d'unité d'ARNr 5.8S, 18S, 28S	3	5	1 #	22 *	100-200	200-400

(†) : en excluant les introns. (‡) : estimation.
(*) : chez *E. cuniculi*, ce sont des ARNr 16S et 23S. Il en existe deux par chromosome.
(#) : chez *D. discoideum*, ce sont des ARNr 5.8S, 17S et 26S. Les deux précurseurs des ARNr sont situés sur un ADN extrachromosomal palindromique de 88 kb présent à ~100 copies par noyau.
Les données concernant les ARNt et les ARNr de *P. falciparum* sont issues de la version 5.0 Bêta de PlasmoDB et ne correspondent pas tout à fait à ce qui a été publié en 2002 [132].
Sources des données pour les autres organismes : *Cryptosporidium parvum* [1], *Encephalitozoon cuniculi* [198], *Dictyostelium discoideum* [101, 361], *Saccharomyces cerevisiae* [407], *Schizosaccharomyces pombe* [406].

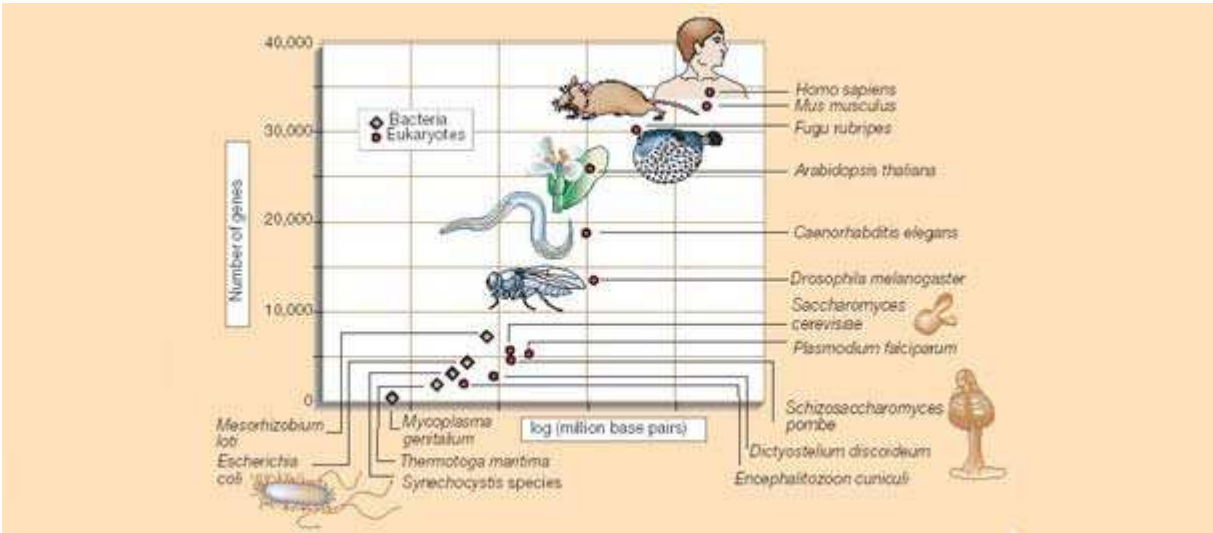


Figure 4. Quelques-uns des génomes entièrement séquencés (2002).
La figure montre le nombre de gènes en fonction de la taille des génomes eucaryotes (●) et procaryotes (◆). A noter, l'échelle logarithmique utilisée pour le taille des génomes, exprimée en million de paires de bases. D'après R.F. Doolittle [93].

La traduction des protéines requiert les trois types d'ARN dans une cellule (Figure 5) :

- l'ARN messager (ARNm) qui porte une partie de l'information génétique contenue au niveau de l'ADN jusqu'au ribosome pour qu'elle soit traduite en protéine ;
- les ARN de transfert (ARNt) qui sont la clé du code génétique : ils transfèrent les acides aminés qui se trouvent dans le cytoplasme jusqu'au ribosome, lieu de la synthèse protéique.
- les ARN ribosomiques (ARNr) qui font partie intégrante des ribosomes. La biogenèse du ribosome, chez les eucaryotes, commence par la synthèse de deux précurseurs d'ARNr par deux ARN polymérases différentes qui seront maturés pour l'un en ARNr 5S et pour l'autre en ARNr 18S, 5.8S et 28S – on parlera alors d'unités 18S-5.8S-28S. Ces ARNr interagiront ensuite avec de nombreuses protéines ribosomiques pour former la petite sous-unité (avec l'ARNr 18S) et la grande sous-unité du ribosome (avec les ARNr 5S, 5.8S et 28S). Lorsque les deux sous-unités s'assemblent, il se forme, entre les deux, un sillon dans lequel passera l'ARNm [127] ;

En règle générale, ce sont les ARNr qui sont de loin les ARN les plus abondants dans une cellule (~82%) alors que les ARNt ne représentent qu'environ 16% et les ARNm environ 2% des ARN totaux d'une cellule.

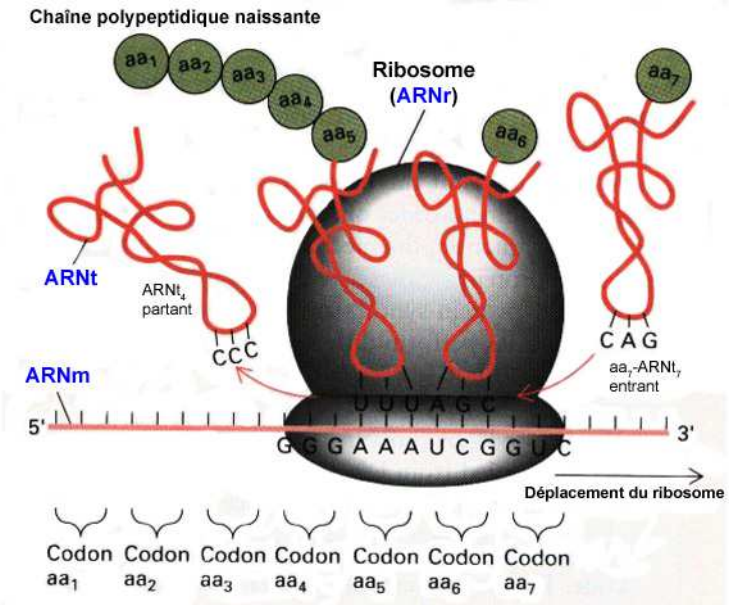


Figure 5. Les trois types d'ARN jouent un rôle dans la synthèse des protéines.
Les trois types d'ARN sont indiqués en bleu. L'ARNm est traduit en protéine par l'action concertée de l'ARNt et du ribosome, qui est une structure hybride composée de nombreuses protéines et de 4 ARNr. Adaptée à partir du livre *An introduction to genetic analysis* [150].

Contrairement à d'autres eucaryotes, *P. falciparum* ne contient pas dans son génome de longues séries de gènes d'ARNr répétés en tandem (Tableau 3). Le parasite possède uniquement 3 unités 18S-5.8S-28S distribuées sur différents chromosomes qui sont toutes différentes les unes des autres au niveau de leur séquence mais aussi au niveau du stade parasitaire où elles sont exprimées [393]. Et l'ARN 5S est codé par trois gènes identiques répétés en tandem sur le chromosome 14.

Chez *P. falciparum*, 45 ARNt ont été identifiés (Tableau 3) : ils représentent tous les acides aminés. Chaque anticodon n'apparaît qu'une seule fois sauf celui de la méthionine pour laquelle un ARNt est réservé à l'initiation de la traduction et un autre code les méthionines internes. De plus, il existerait un anticodon pour la sélénocystéine. Chez la plupart des organismes séquencés jusqu'à aujourd'hui, les gènes codant les ARNt sont assez redondants, les deux seules exceptions connues étant celles des parasites intracellulaires *Cryptosporidium parvum* et *Encephalitozoon cuniculi* qui ne contiennent respectivement que 45 et 44 gènes d'ARNt dans leur génome [198]. Une troisième exception existe maintenant avec *P. falciparum* qui montre une redondance d'ARNt des plus minimales dans son génome nucléaire.

En résumé, le génome de *Plasmodium falciparum* se trouve parmi les plus petits génomes eucaryotes connus à ce jour. Il présente des similitudes avec d'autres génomes entièrement séquencés comme celui d'*Encephalitozoon cuniculi*, mais il montre aussi des particularités comme sa richesse en A+T. En effet, *Plasmodium* est l'organisme qui contient le pourcentage en A+T le plus élevé connu à ce jour.

Le protéome

Des 5 268 protéines prédites, 3 208 sont uniques : elles ne présentent pas assez d'homologie au niveau de leurs séquences avec des protéines d'autres organismes pour que l'on puisse les intégrer dans une famille de gènes et/ou leur assigner une fonction putative. Cela représente environ 60% des protéines, ce qui est beaucoup plus que ce que l'on peut rencontrer chez d'autres organismes [132]. Ceci peut être le reflet de la distance, en terme d'évolution, existant entre *P. falciparum* et les autres eucaryotes dont le génome a été

séquencé, cette distance pouvant être exacerbée par l'extrême richesse en (A+T) du génome du parasite.

Parmi ces 5 268 protéines prédites, 5% ont une similitude significative avec des protéines malheureusement encore hypothétiques existant chez d'autres organismes. Un tiers des protéines prédites présentent un ou plusieurs domaines transmembranaires et 17,3% possèdent un peptide signal ou un signal d'ancrage putatif.

La base de données Gene Ontology (GO) [14] fournit un vocabulaire très structuré pour des domaines biologiques spécifiques permettant de décrire des produits de gènes dans un organisme donné. Des termes de GO ont été assignés manuellement à 2 134 produits de gènes (~40%) de *P. falciparum*. Quand on compare cette annotation avec l'annotation des produits des gènes de *S. cerevisiae* [132], on se rend compte qu'on a pu appliquer des termes GO à une plus grande proportion des produits de gènes de la levure et ce, que l'on utilise le principe d'organisation selon la fonction moléculaire ou le processus biologique. Ceci vient du fait que le génome de la levure est bien mieux caractérisé que celui du parasite. Cependant, il existe deux exceptions, reflétant deux processus typiques du parasite : la catégorie 'cell invasion or adhesion' et la catégorie 'physiological processes' dans laquelle on retrouve les 208 gènes connus pour être impliqués dans l'échappement du système immunitaire de l'hôte.

Comme il existe des processus typiques du parasite, il existe des processus typiques de la levure ('sporulation' et 'cell budding'). Néanmoins, **très peu de produits de gènes de *P. falciparum* sont associés aux catégories suivantes : 'cell organization and biogenesis', 'cell cycle' ou 'transcription factor', par comparaison avec *S. cerevisiae*. Ces différences n'impliquent pas nécessairement qu'un plus faible nombre de gènes parasitaires soit impliqué dans ces processus, mais soulignent les domaines de la biologie du parasite où les connaissances sont encore limitées**.

LA TRANSCRIPTION ET LA REGULATION TRANSCRIPTIONNELLE CHEZ LES EUCARYOTES

Chaque animal est constitué d'un ensemble d'unités vivantes qui portent chacune en soi toutes les caractéristiques de la vie. Rudolph Virchow (1858)

Les êtres vivants sont faits d'éléments si étroitement interdépendants que l'on ne juge bien de l'importance d'aucun d'eux sans tenir compte des autres. Harvey Lodish [238]

Ce que Rudolph Virchow a dit des animaux peut être appliqué à tout être vivant. En effet, tous les êtres vivants ont comme point commun d'être composé d'un élément de base : la cellule. La compréhension de la complexité d'une cellule passe par l'étude des nombreux éléments interdépendants qui la composent et nécessite de nombreuses disciplines de la biologie.

Dans la cellule se trouvent, entre autres, les gènes, c'est-à-dire les éléments qui déterminent non seulement la structure des protéines, mais aussi l'infrastructure cellulaire et orchestrent tout ce qui est nécessaire à la cohésion d'un organisme, qu'il soit uni- ou pluricellulaire. Dans ce chapitre sera abordé une des étapes de l'expression des gènes : la transcription et les mécanismes qui permettent de la moduler. Nous verrons que même si cette étape est une petite partie de la régulation de l'expression de gènes et de ce qui permet à une cellule de vivre, elle est extraordinairement complexe.

I - Structure du matériel génétique chez les eucaryotes

L'existence d'un noyau vrai contenant l'ADN génomique en tant qu'organite délimité par une enveloppe membranaire est une caractéristique de la cellule eucaryote. C'est un organite permanent de la vie de la cellule dite « interphasique ». Pendant l'interphase s'effectuent dans le noyau :

- la transcription des messages codés dans l'ADN qui interviennent dans la prolifération et/ou la différenciation des cellules ainsi que dans leurs activités physiologiques,
- la préparation à la division cellulaire grâce à la réplication de l'ADN dans les cellules appelées à se diviser.

I.1 - Structure de la chromatine

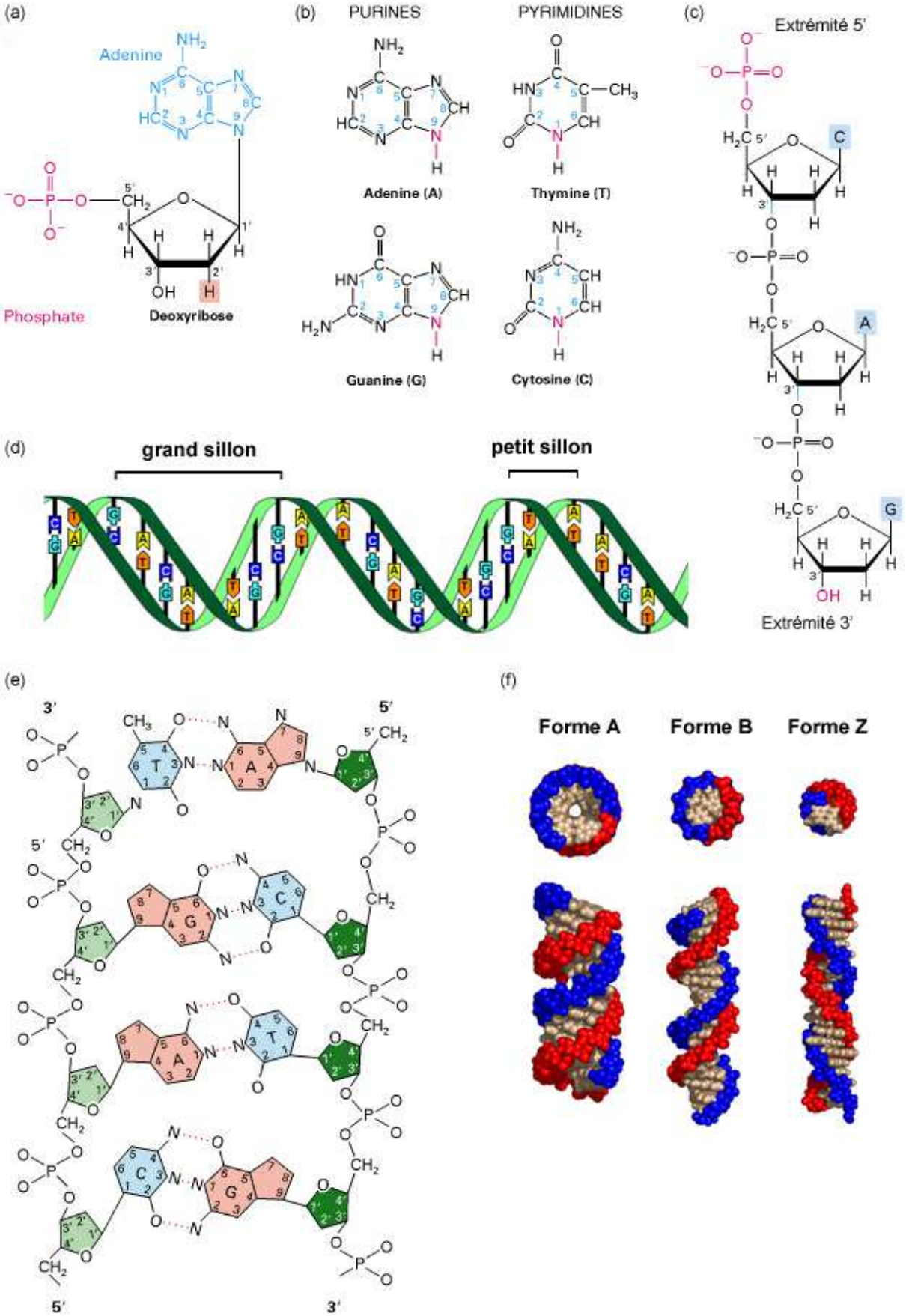
L'ADN est le patrimoine génétique qui contient l'information nécessaire à la construction d'une cellule ou d'un organisme. La tâche monumentale accomplie par la biologie moléculaire depuis la découverte en 1953 de la structure de l'ADN par James Watson et Francis Crick (Prix Nobel de médecine et physiologie en 1962) [395], jusqu'à la fin des années 1970 a élucidé en profondeur la structure de l'ADN ainsi que la synthèse d'ADN, d'ARN et des protéines.

L'ADN est formé de quatre types de bases, appelées nucléotides, formées de trois parties : un groupe phosphate, un pentose (sucre à cinq carbones) et une base organique (Figure 6a). Dans l'ADN, le pentose est toujours le désoxyribose et les quatre bases organiques sont l'adénine et la guanine (les purines) ainsi que la thymine et la cytosine (les pyrimidines), généralement représentées par les sigles A, G, T et C (Figure 6b). Lorsque les nucléotides se polymérisent en acides nucléiques, des liaisons phosphodiester se créent entre le sucre du premier nucléotide et le groupe phosphate du deuxième (Figure 6c). Comme tout polypeptide, un brin d'acide nucléique a une orientation chimique : l'extrémité 5′ porte un groupe ester-phosphate sur le carbone 5′ du sucre et l'extrémité 3′ porte un groupe hydroxyle libre sur le carbone 3′ du sucre. Cette polarité, et qui plus est, le fait que la synthèse procède de 5′ en 3′, est à la base de la convention selon laquelle les polynucléotides sont tracés et lus dans le sens 5′ → 3′.

L'ADN comporte deux chaînes polynucléotidiques : les brins, torsadées l'une autour de l'autre en forme de double hélice ou hélice bicaténaire (Figure 6d). A l'extérieur courent les deux squelettes sucre-phosphate, les bases étant tournées vers l'intérieur. Dans un brin, les bases adjacentes s'empilent en plans parallèles. L'orientation des brins est anti-parallèle, ils sont maintenus en contact pas des liaisons hydrogène et des interactions hydrophobes. Les bases appartenant à des brins opposés sont toujours alignées avec précision par appariement strict d'une base d'un brin à une base de l'autre brin : A s'apparie avec T par deux liaisons hydrogène et G s'apparie avec C par trois liaisons hydrogène (Figure 6e). La complémentarité des bases résulte de la taille, de la forme et de la composition chimique des bases. Des interactions hydrophobes et de van der Waals entre les paires de bases adjacentes dans la pile contribuent d'une façon significative à la stabilité globale de la double hélice.

L'ADN existe sous plusieurs formes (Figure 6f). La forme B est la forme biologique la plus importante et la plus courante. Elle correspond à la forme décrite en 1953 par Watson & Crick [395]. Les bases empilées sont régulièrement espacées de 3,4 nm le long de l'axe de l'hélice de pas droit. En parcourant un tour complet de l'hélice, on avance de 3,4 nm, d'où la présence d'environ 10 paires de bases (pb) par tour. A l'extérieur de la molécule de forme B, les brins torsadés ménagent entre eux deux sillons hélicoïdaux de largeur différente : le grand sillon (1,2 nm de large) et le petit sillon (0,6 nm de large). Il existe une forme A d'ADN qui compte 11 bases par tour d'hélice, lequel couvre une distance de 2,3 nm. L'ADN prend sa forme A dans les solutions non aqueuses. Certains courts segments d'ADN peuvent adopter, au lieu de la configuration droite normale, une configuration de pas gauche appelée ADN-Z. La forme Z de l'ADN a été initialement une création de laboratoire avec l'oligonucléotide artificiel d(CGCGCG) [389], dans des conditions physico-chimiques qui n'avaient rien de conditions physiologiques. Aujourd'hui, l'ADN-Z a été décelé dans des chromosomes de mammifères, mais sa fonction précise est encore mal connue.

Figure 6. Structure, appariement et formes de l'ADN (→).
(a) Structure chimique d'un nucléotide typique. Un nucléotide est constitué d'un groupement phosphate (en rouge), d'un pentose (en noir) qui pour l'ADN est le désoxyribose et d'une base (en bleu), ici l'adénine. **(b)** Structure chimique des bases constituant les nucléotides. Les purines (A et G) sont constituées de deux cycles tandis que les pyrimidines (C et T) n'en contiennent qu'un. L'azote 9 des purines et l'azote 1 des pyrimidines (en rouge) sont liés au carbone 1′ du désoxyribose. **(c)** Les nucléotides se polymérisent en acides nucléiques grâce à la formation de liaisons phosphodiester (en bleu). Le nucléotide de l'extrémité 5′ porte un groupe phosphoryle libre, et celui de l'extrémité 3′ un désoxyribose à hydroxyle 3′ libre. **(d)** Modèle de la double hélice d'ADN. Le squelette sucre-phosphate est à l'extérieur et les bases sont tournées vers l'intérieur, empilées en plateaux parallèles. L'ADN présente deux sillons de tailles différentes : le petit et le grand sillons. **(e)** Structure déployée d'un double brin d'ADN pour montrer les squelettes sucre-phosphate (les sucres en vert et les phosphates en noir), les bases appariées et les liaisons hydrogènes qui les unissent (en pointillés rouges). Les chaînes sont antiparallèles. Chaque paire de bases consiste en une purine (A ou G, en rose) et une pyrimidine (T ou C, en bleu clair) unies par des liaisons hydrogènes. **(f)** Vues transversales et longitudinales des 3 formes de l'ADN. Sont représentés en rouge et bleu le squelette sucre phosphate de chaque brin d'ADN et en beige clair les bases appariées. Les parties (a), (b) et (c) sont tirées du livre *Molecular cell biology* [238] et la partie (e) de *The DNA helix and how it is read* [89].


(a)
Adenine
Phosphate
Deoxyribose
(b)
PURINES
PYRIMIDINES
Adenine (A)
Thymine (T)
Guanine (G)
Cytosine (C)
(c)
Extrémité 5'
Extrémité 3'
(d)
grand sillon
petit sillon
(e)
3'
5'
(f)
Forme A
Forme B
Forme Z

La longueur totale de l'ADN atteignant plusieurs centaines de milliers de fois celle de la cellule, la compression de l'ADN en chromosome est indispensable à l'architecture de la cellule (Figure 7a). L'ADN est alors associé à des protéines pour former la chromatine. Les fibres chromatiniennes, obtenues par extraction, possèdent une structure en « collier de perles », une perle correspondant à un **nucléosome**. Chaque nucléosome est constitué par un cœur protéique formé par les **histones H2A, H2B, H3 et H4** organisées en un octamère (4 paires, une paire par type d'histone) sur lequel s'enroule l'ADN à peu près deux fois et chaque nucléosome est séparé du suivant par un ADN nu, appelé ADN de liaison, dont la taille varie selon les espèces et le type cellulaire. Ce module de base se répète régulièrement pour former un nucléofilament qui peut s'organiser lui-même en structures de plus en plus compactes. Un cinquième type d'histone, l'**histone H1**, vient s'associer au nucléosome pour le stabiliser. On parle alors de forme condensée de l'ADN (ø : 11 nm). Cette forme peut se spiraliser et former des solénoïdes rendus solidaires par les histones H1 (Figure 7b). Celles-ci se trouvent alors situées dans la face interne du solénoïde (ø : 30 nm). Une spiralisation de deuxième ordre donne des nucléofilaments de 100 nm de diamètre qui peuvent ensuite s'enrouler sur eux-mêmes et former des filaments de 300 nm. Puis la spiralisation de ces filaments constituent des nucléofilaments de 700 nm. Au moment de la mitose, les fibres chromatiniennes s'organisent en chromosomes.

La structure du nucléosome n'est pas figée : elle est influencée par de nombreux facteurs qui peuvent la modifier. Cette structure dynamique est déterminante pour toutes les fonctions du génome : la recombinaison, la réplication, la réparation et la transcription.

La chromatine des régions chromosomiques qui ne sont pas transcrites existe en majeure partie sous la forme condensée (hétérochromatine), alors que les régions en train d'être transcrites prennent une forme décondensée. En effet, juste avant la transcription, la structure de la fibre chromatinienne se modifie : le gène qui doit être transcrit se déroule, les histones H1 se détachent temporairement, le temps de la transcription. La fibre chromatinienne est localement décondensée et cette décondensation s'effectue notamment grâce à des **protéines non-histones** dont je parlerai ultérieurement.

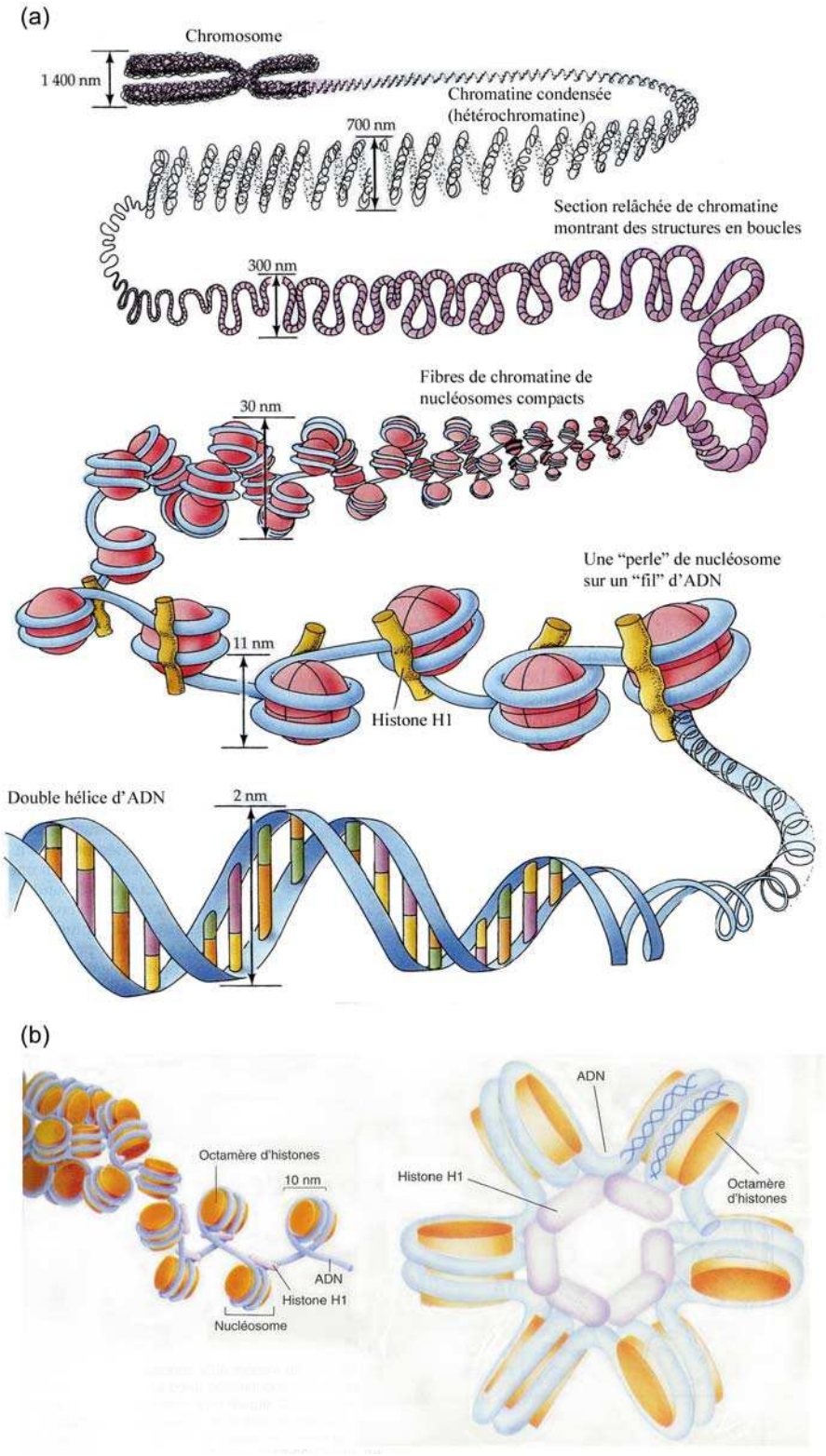


Figure 7. Niveaux d'empaquetage de l'ADN dans un chromosome.
(a) Schéma montrant comment l'ADN est « empaqueté » dans un chromosome. Adaptée du livre *Life : the science of biology* [312]. **(b)** Deux aspects d'un modèle de fibre chromatinienne condensée. Le cœur octamérique d'histones y a la forme d'un disque. A chaque nucléosome s'attache une histone H1 et la fibre se tord pour former une structure solénoïdale de 30 nm de diamètre. Adaptée du livre *Molecular cell biology* [238].

I.2 - Définitions d'un gène

En termes moléculaires, un gène est défini comme l'ensemble de la séquence nucléotidique indispensable à la synthèse d'une molécule d'ARN fonctionnel [238]. Selon cette définition, un gène n'englobe pas seulement les nucléotides spécifiant soit une protéine, c'est-à-dire la région codante ou unité de transcription, soit un ARN fonctionnel, tel un ARNr ou un ARNt, mais aussi toutes les séquences d'ADN nécessaires à la formation du transcrit primaire. Dans les gènes eucaryotes, on peut trouver certaines séquences d'ADN régulant l'amorçage de la transcription par l'ARN polymérase jusqu'à au moins 50 kilobases de la région codante. Mais, dans ce manuscrit, je considèrerai qu'un gène correspond à l'unité de transcription uniquement et toutes les séquences nécessaires à la formation du transcrit primaire forment le promoteur.

Chez les eucaryotes, contrairement à ce que l'on rencontre chez les procaryotes, les gènes (Figure 8) sont monocistroniques, c'est-à-dire qu'ils sont physiquement séparés les uns des autres sur l'ADN et sont transcrits à partir d'un site d'initiation qui leur est propre pour la synthèse de leur ARN. Les gènes sont morcelés en séquences codantes (en violet foncé sur la Figure 8) et séquences non codantes (en jaune et en violet clair sur la Figure 8) : on parle de gènes « en mosaïque ». Chaque gène débute, par définition, par un site d'initiation de la transcription, arbitrairement désigné comme étant le point +1, et se termine après le signal de fin de transcription ou signal de polyadénylation. Deux régions considérées comme des exons, qui sont donc transcrites mais qui ne seront pas traduites, encadrent la phase ouverte de lecture : la région 5'UTR (untranslated region) permet aux ribosomes d'accrocher l'ARNm afin que la traduction débute à l'AUG initial et la région 3'UTR contient le signal de polyadénylation qui permettra à une enzyme particulière d'ajouter une queue poly-A lors de la maturation du transcrit (voir p. 52).

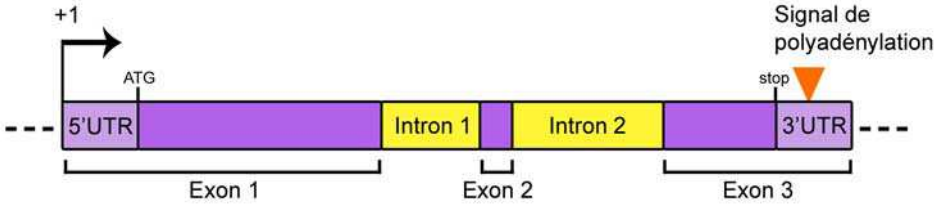


Figure 8. Structure d'une unité de transcription.
Une unité de transcription est constituée de séquences codantes (en violet foncé) et de séquences non codantes (en jaune). Des régions qui seront transcrites mais non traduites (5' et 3' UTR en violet clair) entourent la phase ouverte de lecture délimitée par l'ATG initiateur et le codon stop.

I.3 - Structure du promoteur

Chez les eucaryotes, il existe de nombreuses « définitions » du promoteur. L'une d'entre elles consiste à considérer qu'il se divise en deux parties : le promoteur proximal et le promoteur distal (Figure 9).

Ces deux parties se définissent essentiellement en fonction de la nature des motifs ou éléments de régulation qu'ils comportent. Les motifs dits « **constitutifs** » se situent dans le promoteur proximal et se retrouvent dans la plupart des promoteurs [43]. Le promoteur distal, quant à lui, comporte des éléments qui sont **spécifiques** d'un gène donné ou d'un groupe de gènes donnés (même si on peut retrouver ces motifs dits « spécifiques » dans le promoteur proximal, voire dans l'unité de transcription).

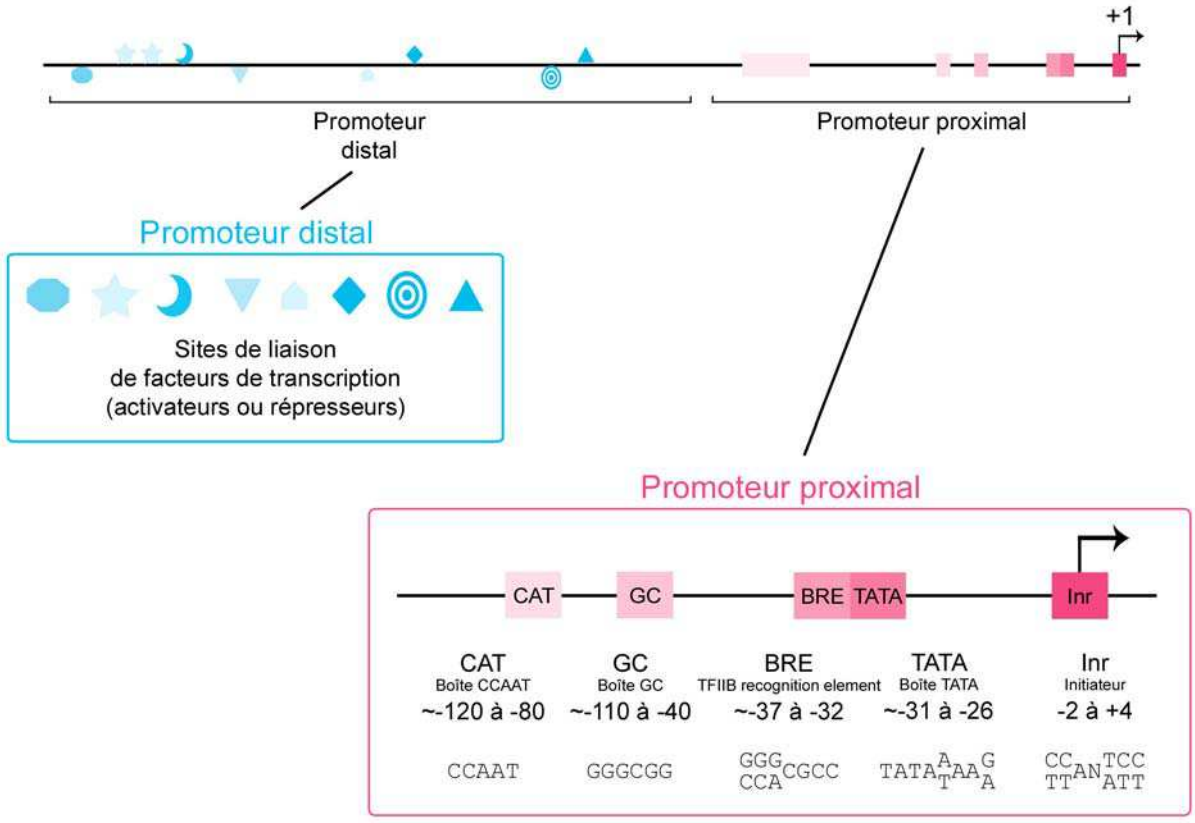


Figure 9. Structure du promoteur.
Le promoteur proximal contient des motifs dits « constitutifs » c'est-à-dire que l'on retrouve dans la plupart des promoteurs de gènes de classe II et sur lesquels se fixent la machinerie basale de transcription et quelques co-facteurs. Le promoteur distal comporte des motifs dits « spécifiques » qui sont les sites de fixation des facteurs de transcription qui permettront de moduler l'expression du gène. Les positions et consensus des différents motifs sont issus de la revue de Smale & Kadonaga [347] et de l'*Abrégé de Biochimie génétique, Biochimie moléculaire* de Jacqueline Etienne [107].

Le promoteur proximal

Juste en amont de l'unité de transcription se situe la région que l'on appelle le **promoteur proximal** : cette séquence d'ADN minimale est suffisante pour promouvoir une transcription de base. La plupart des éléments de cette région interagissent avec des composants de la **machinerie basale de transcription** (voir p. 50). Ce promoteur proximal contient différentes séquences importantes [347], dont le site d'initiation de la transcription. Une séquence riche en A+T, appelée la **boîte TATA** se situe en amont du site d'initiation de la transcription, à -25/-30 [2] chez les eucaryotes supérieurs et de -40 à -120 chez la levure [359]. La boîte TATA est le site de fixation de la « TATA-binding protein » (TBP), une des protéines composant la machinerie basale de transcription. Bien qu'une séquence canonique puisse être déduite de toutes les boîtes TATA connues aujourd'hui, la TBP peut se lier et fonctionner avec une large gamme de séquences, ce qui rend difficile l'identification d'une boîte TATA sur un promoteur seulement avec une séquence. Dans certains gènes, le site d'initiation de la transcription fait partie d'un élément **Initiateur** (Inr) sur lequel peuvent se fixer divers facteurs, ce qui facilite le recrutement de la machinerie basale de transcription.

Un promoteur proximal peut contenir soit une boîte TATA et un Initiateur, soit un seul de ces éléments, soit aucun. Il est connu que la plupart des gènes de classe II, c'est-à-dire les gènes transcrits par l'ARN polymérase II et donc les gènes auxquels je me suis intéressée chez *Plasmodium*, présente des promoteurs contenant uniquement une boîte TATA et très peu ont seulement un Initiateur.

Le promoteur proximal peut aussi comporter la **séquence BRE** (TFIIB-recognition element) située juste en amont de la boîte TATA et reconnue par le facteur TFIIB, un autre facteur de la machinerie basale de transcription ainsi que deux boîtes caractéristiques :

- la **boîte GC** qui serait reconnue uniquement par le facteur de transcription Sp1,
- la **boîte CCAAT** sur laquelle seraient capables de se fixer, entre autres, les facteurs NF-Y (nuclear factor-Y aussi appelé CBF pour CCAAT-box-binding factor) et NF-I (nuclear factor-I aussi appelé CTF pour CCAAT-binding transcription factor).

[2] La numérotation des bases se fait en fonction du point +1, la base se situant juste avant le point +1 est numérotée -1.

Le promoteur distal

Bien que les éléments dits « constitutifs » du promoteur proximal soient fondamentaux pour le recrutement de la machinerie basale de transcription, la composition et le contexte des éléments dits « spécifiques » peuvent influencer la régulation transcriptionnelle.

Au départ, les chercheurs se sont particulièrement intéressés aux séquences correspondant aux régions situées immédiatement en amont de l'extrémité 5′ des phases codantes, c'est-à-dire le promoteur proximal, car ils pensaient qu'elles régulaient la synthèse des ARNm chez les eucaryotes. Ceci est en partie vrai mais il existe aussi toute une série d'autres séquences d'ADN dites « spécifiques », généralement constituées de quelques nucléotides et dispersées tout au long de la région située en 5′ du gène (mais pouvant également se trouver en aval du point +1) que l'on appelle des **éléments *cis*-régulateurs**. Ces éléments, selon leur position par rapport au site d'initiation de la transcription, font partie du promoteur proximal ou du **promoteur distal**. Sur ces séquences précises d'ADN viennent se fixer des **facteurs *trans*-régulateurs** ou **facteurs de transcription**. Ces facteurs sont indispensables à la transcription car ce sont eux qui régulent le niveau de transcription. En effet, le niveau maximal auquel peut être transcrit un gène déterminé dépend de la séquence nucléotidique du promoteur dans son ensemble et, d'un gène à un autre, ce niveau peut varier d'un facteur 1000.

II - La transcription

II.1 - L'expression des gènes codant des protéines

Le flux de l'information cellulaire passe notamment par l'expression de certains gènes d'une cellule. Ces gènes sont localisés sur l'ADN dans le noyau et leurs produits finaux se situent dans le cytoplasme sous forme de protéines. Il existe donc un intermédiaire entre l'ADN et les protéines : l'ARN messager. L'ARNm, comme les ARNr et les ARNt, est synthétisé dans le sens 5′ → 3′ de manière antiparallèle par rapport au brin d'ADN transcrit et complémentaire (Figure 10). Il copie donc les messages codés dans la molécule d'ADN et dirige la synthèse des protéines dont certaines catalysent la polymérisation même de l'ADN

et de l'ARN. Ainsi, l'ADN dirige la synthèse de l'ARN, ce qui assure la **transcription** de l'information, et l'ARN dirige ensuite la synthèse de la protéine, ce qui correspond à la **traduction** de l'information.

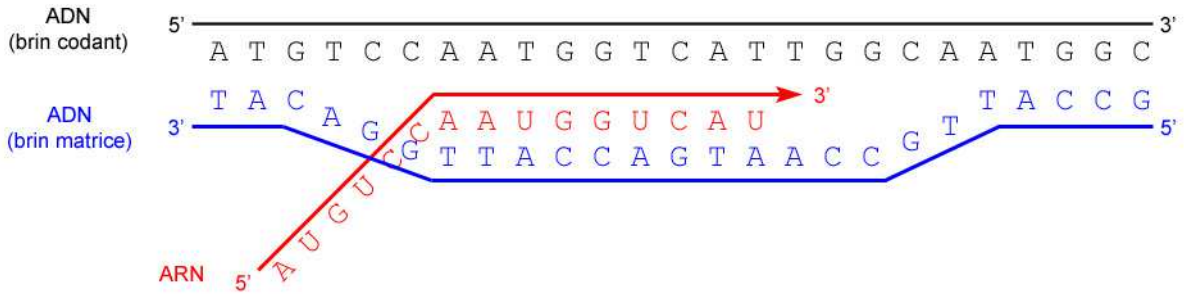


Figure 10. Synthèse d'un ARN par complémentarité d'un brin d'ADN de la double hélice.
Les deux brins d'ADN sont écartés par rupture des liaisons hydrogène de façon à ce que l'ARN polymérase (non représentée ici) puisse synthétiser un ARN (en rouge) par complémentarité du brin qui doit être transcrit : le brin matrice (en bleu). L'ARN a donc la même polarité et la même séquence en bases (à l'exception de T remplacé par U) que le brin codant (en noir). La flèche indique le sens de la synthèse.

Les étapes conduisant de l'ADN aux protéines sont nombreuses et toutes doivent être contrôlées et régulées par la cellule. En effet, la capacité à exprimer une protéine biologiquement dépend de différents « points de contrôle » entre régulations transcriptionnelle et post-transcriptionnelle :

- la **structure de la chromatine** : la structure physique de l'ADN peut affecter la capacité des facteurs de transcription et des ARN polymérases à accéder à des gènes spécifiques et à activer leur transcription ;
- l'**initiation de la transcription** : c'est l'étape la plus importante pour le contrôle de l'expression des gènes. La présence ou l'absence ainsi que la force des sites de liaison des facteurs de transcription et l'interaction entre les facteurs, qu'ils soient activateurs ou inhibiteurs, sont autant d'éléments qui régissent l'initiation de la transcription ;
- la maturation du transcrit : les ARNm eucaryotes doivent être coiffés et polyadénylés et les introns doivent être éliminés. Certains gènes peuvent aussi subir un épissage alternatif, ce qui fait qu'à partir d'un seul et unique gène, on obtient deux ou plus de deux protéines différentes ;

- le transport de l'ARN hors du noyau : les ARN, qui doivent être correctement maturés, sont pris en charge par des protéines qui les amènent du noyau jusqu'aux ribosomes dans le cytoplasme ;
- la stabilité du transcrit : la stabilité des transcrits varie énormément d'un transcrit à l'autre ;
- l'initiation de la traduction : puisque les ARNm possèdent plusieurs codons AUG, la capacité des ribosomes à reconnaître et à initier la traduction au niveau du bon codon peut affecter l'expression du produit d'un gène ;
- les modifications post-traductionnelles : des modifications communes, comme la glycosylation, l'acétylation, la phosphorylation ou encore la formation de ponts disulfures sont indispensables au bon fonctionnement des protéines ;
- le transport de la protéine : une protéine peut-être nucléaire, cytoplasmique, membranaire ou encore exogène et à chaque destination correspond un moyen de transport ;
- la stabilité de la protéine : certaines protéines sont très stables alors que d'autres le sont très peu.

Bien que toutes ces étapes soient essentielles à l'expression d'un gène, je me focaliserai essentiellement sur les deux premières étapes, c'est-à-dire la structure de la chromatine et l'initiation de la transcription.

II.2 - Les ARN polymérases

Les noyaux de toutes les cellules eucaryotes étudiées jusqu'à présent contiennent trois ARN polymérases différentes, désignées I, II et III [325]. Chaque ARN polymérase catalyse la transcription de gènes codant pour une classe particulière d'ARN. L'**ARN polymérase I** est située dans le nucléole et sert à la synthèse du précurseur d'ARNr dont seront issus ensuite les ARNr 18S, 5.8S et 28S qui entrent dans la composition du ribosome. L'**ARN polymérase III** fonctionne en dehors du nucléole et transcrit les gènes d'ARNt, d'ARN 5S et de beaucoup d'autres petits ARN stables. Quant à l'**ARN polymérase II**, elle catalyse la transcription de tous les gènes codant des protéines ; on lui doit donc la formation de tous les ARNm. Elle produit aussi quatre petits ARN intervenant dans l'excision-épissage de l'ARN.

Je m'intéresserai par la suite uniquement à la transcription par l'ARN polymérase II.

II.3 - Le cycle de la transcription par l'ARN polymérase II

La transcription d'un gène par l'ARN polymérase II (Figure 11) commence par la liaison de facteurs qui modifient la structure de la chromatine ce qui facilite la fixation de la **machinerie basale de transcription** au niveau du promoteur proximal [67, 310]. L'ARN polymérase II a besoin de facteurs d'amorçage présents dans le noyau de la cellule pour pouvoir reconnaître les sites d'initiation de la transcription [248, 399]. Ainsi, plusieurs protéines s'assemblent de concert avec l'ARN polymérase II au niveau du promoteur proximal en un complexe d'amorçage de la transcription presque aussi volumineux que le ribosome : le **complexe de pré-initiation** ou **complexe fermé**. On a donné à ces protéines le nom de **facteurs généraux de la transcription** parce qu'elles semblent indispensables à la transcription de tous les gènes transcrits par cette enzyme, contrairement aux facteurs de transcription (qui seront décrits après) qui s'attachent à des sites précis et spécifiques que l'on ne trouve que dans les promoteurs de certains gènes. Ces facteurs généraux rejoignent le promoteur proximal dans un ordre bien établi (Figure 11) [45, 65, 416].

Le premier facteur à rejoindre le promoteur proximal est le facteur général **TFIID** [80, 117]. Ce facteur est en fait un complexe constitué de plusieurs protomères : la TBP qui se fixe sur la boîte TATA [272] sans grande spécificité d'orientation [69] est associée à des co-facteurs que l'on nomme les TAF (TBP-associated factors) [311, 370] dont certains se lient à l'Inr [56, 283]. La TBP, en se fixant à l'ADN dans son petit sillon, impose une courbure considérable à l'ADN [203].

Des inhibiteurs peuvent se joindre au complexe ADN-TFIID et empêcher d'autres facteurs de s'y joindre à moins que le facteur **TFIIA** ne soit venu s'y attacher en interagissant directement avec la TBP [138, 368], permettant ainsi le maintien du complexe déjà établi [45, 398] et la poursuite du processus d'assemblage [44].

Le complexe est alors rejoint par le facteur **TFIIB** [321, 418], qui reconnaît la TBP ainsi que la distorsion de l'ADN [280] et se fixe sur la séquence BRE avec une grande affinité. Ce facteur est indispensable pour le recrutement, ensuite, de l'ensemble **TFIIF-ARN polymérase II** [45, 119, 351]. Cependant, la polymérase n'initiera la transcription qu'après

fixation de deux autres facteurs, à savoir dans l'ordre les facteurs **TFIIE** [321] et **TFIIH** [63, 367], TFIIE étant impliqué dans le recrutement de TFIIH [45, 120].

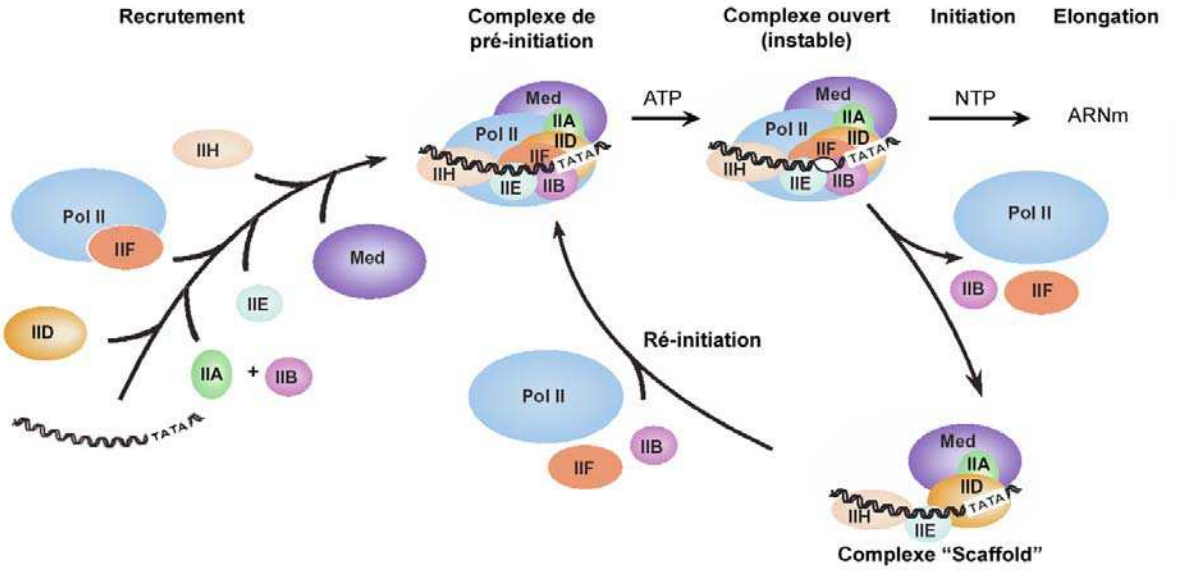


Figure 11. Assemblage du complexe de pré-initiation sur un promoteur comportant une boîte TATA, suivi de l'initiation et de la ré-initiation de la transcription par l'ARN polymérase II.
Les facteurs généraux de la transcription sont nommés TFIIA, TFIIB, etc. : TF pour « transcription factor », II parce qu'ils sont spécifiques de l'ARN polymérase II et une lettre pour les identifier. Le médiateur (Med) peut être recruté dans le complexe de pré-initiation à tout moment [228]. L'ATP est nécessaire pour transformer le complexe fermé en complexe ouvert qui, grâce aux nucléosides triphosphates (NTP), pourra synthétiser un ARNm. Adapté de l'article de S. Hahn [156].

Bien que l'ARN polymérase II et les facteurs généraux de la transcription soient suffisants pour initier une transcription, ce système ne peut répondre aux facteurs de transcription qu'en présence d'un autre complexe protéique appelé **médiateur** [228]. Le médiateur interagit avec le domaine C-terminal de la plus grande sous-unité de l'ARN polymérase II et avec certains facteurs généraux.

Le promoteur proximal sert à positionner le complexe de pré-initiation au bon endroit et dans le bon sens. Dans cet état, la polymérase et les facteurs généraux sont tous liés au promoteur mais ne sont pas dans une conformation active pour commencer la transcription. Un changement conformationnel intervient : le facteur TFIIH phosphoryle le domaine C-terminal de l'ARN polymérase II grâce à son activité kinase et se sert de l'énergie libérée par cette phosphorylation pour séparer les deux brins d'ADN grâce à son activité hélicase. Le brin d'ADN qui doit être transcrit, c'est-à-dire le brin matrice (Figure 10), se positionne alors dans la « fissure » du site actif de l'ARN polymérase II pour former le **complexe ouvert** [391].

L'initiation de la transcription commence alors avec la synthèse de la première liaison phosphodiester de l'ARNm. Après la synthèse d'environ 30 bases d'ARNm, la polymérase interrompt ses contacts avec le promoteur proximal et le reste de la machinerie de transcription pour entrer dans l'étape d'élongation de la transcription. Elle se déplace le long de l'ADN, en ouvrant une partie de la molécule d'ADN par un mécanisme de déroulement de l'hélice. Les facteurs qui permettent la synthèse d'ARNm « productifs », la maturation et l'export des ARNm ainsi que la modification de la chromatine sont recrutés par l'ARN polymérase au cours de l'élongation [25]. Après l'initiation de la transcription par l'ARN polymérase *in vitro*, beaucoup de facteurs généraux de la transcription restent derrière, au niveau du promoteur proximal [414] dans un complexe, appelé complexe en échafaudage ou complexe « Scaffold », qui marque les gènes ayant déjà été transcrits et qui rend l'étape de recrutement des facteurs généraux manquants beaucoup plus rapide pour une ré-initiation de la transcription.

La terminaison de la transcription est un processus compliqué et critique pour l'expression réussie d'un gène : elle permet la libération du transcrit primaire du site actif de l'ARN polymérase II ainsi que la libération de l'ARN polymérase II de l'ADN. De plus, elle assure qu'un promoteur ne soit pas perturbé par la polymérase transcrivant le gène situé en amont [309] : ceci diminue l'interférence qui peut exister quand deux gènes proches sur la chromatine sont exprimés en même temps.

II.4 - Maturation des transcrits

Avant de former un ARNm utilisable, la chaîne d'ARN précurseur produite par l'ARN polymérase II, appelée transcrit primaire, subit plusieurs étapes de maturation en même temps que la transcription s'effectue (Figure 12). Ces « transformations » du transcrit primaire se font grâce à des régions-clés non codantes dont la mutation empêche la formation d'un ARNm fonctionnel et par extension la formation du polypeptide correspondant.

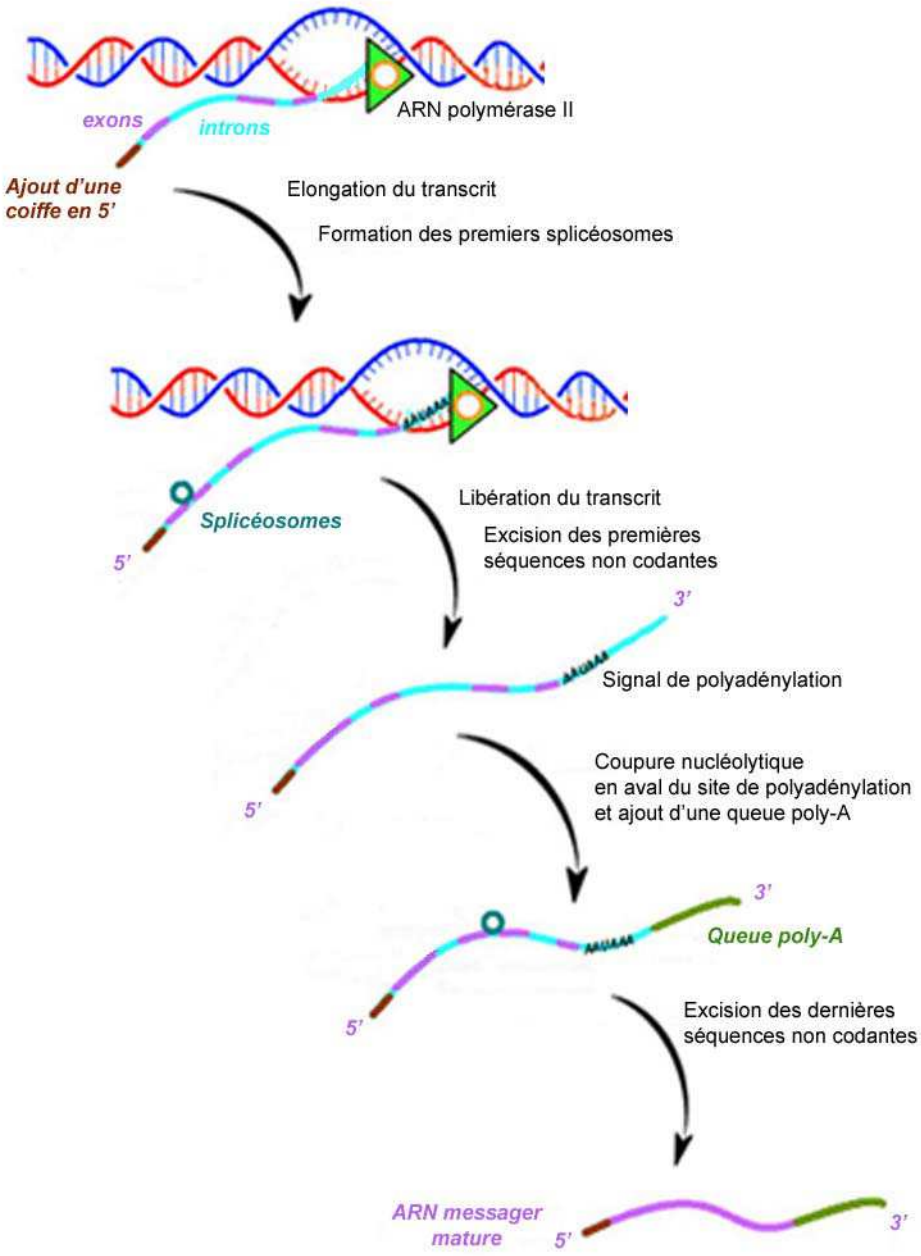


Figure 12. Maturation des transcrits primaires.
Alors que le transcrit primaire est en cours d'élongation, une coiffe est ajoutée en 5' pour le protéger des dégradations enzymatiques et les introns forment alors des boucles d'ADN que l'on appelle, quand elles sont associées avec des ribonucléoprotéines nucléaires (non représentées ici), des splicéosomes. Ces séquences non codantes sont alors excisées au fur et à mesure de leur formation. Une fois le transcrit terminé et libéré de l'ARN polymérase II, une coupure nucléolytique a lieu quelques nucléotides en aval du site de polyadénylation et une queue poly-A est ajoutée à la fin du transcrit. Une fois toutes les séquences non codantes excisées, l'ARNm mature pourra être transféré dans le cytoplasme où il sera traduit. Adaptée à partir d'une image tirée de l'URL http://www.geneticengineering.org/chemis/Chemis-NucleicAcid/ARN.htm.

Au nucléotide situé en 5' vient s'ajouter une coiffe méthylée ; cette structure, ajoutée au transcrit primaire en cours de transcription, joue plusieurs rôles :

- elle protège les ARNm des dégradations enzymatiques ;
- elle fournit également un signal de reconnaissance pour les protéines impliquées dans le processus ultérieur de maturation ;
- une fois que l'ARNm mature est sorti du noyau, elle aide les ribosomes à initier la traduction au codon initiateur AUG correct.

La maturation du transcrit primaire continue par l'élimination les introns. Cette élimination s'effectue, tout au long de l'élongation de la transcription, par l'excision des introns qui forment des boucles d'ARN auxquelles sont associées des particules ribonucléoprotéiques nucléaires jouant le rôle de catalyseurs ; le complexe formé par ces particules fixées aux boucles d'ARNm s'appellent des splicéosomes. L'excision des introns sera alors suivie d'un épissage qui « ressoude » les exons séparés auparavant par les introns. Les introns excisés seront hydrolysés en nucléosides par des exonucléases et ainsi recyclés.

Puis, une fois le transcrit primaire libéré de l'ARN polymérase II, une coupure nucléolytique a lieu à l'extrémité 3' ce qui sectionne le messager en un point situé une vingtaine de bases après le signal de polyadénylation : l'extrémité 3' ainsi formée est alors allongée par une série de résidus adénylate grâce à une enzyme appelée poly-A polymérase. La queue poly-A ainsi formée comprend, selon l'espèce, de 100 à 250 nucléotides et stabilise l'ARNm [309].

L'ARNm maturé est transféré dans le cytoplasme où il pourra être traduit. La découverte de l'épissage des ARNm, et donc de la structure en mosaïque des gènes, a été pour la première fois annoncée en 1977 lors d'un congrès sur la chromatine à Cold Spring Harbor (Etats-Unis) par Richard J. Roberts & Phillip A. Sharp, ce qui leur a valu de recevoir le Prix Nobel de médecine et physiologie en 1993.

III - La régulation de la transcription

La transcription des gènes eucaryotes est précédée de multiples événements : la décondensation du début du locus [3], le remodelage des nucléosomes, les modifications des histones et le recrutement de la machinerie basale de transcription au niveau du promoteur proximal. Ce promoteur proximal est suffisant pour diriger l'initiation de la transcription. Néanmoins, l'ARN polymérase II et les facteurs généraux de la transcription ne peuvent pas reconstituer à eux seuls les niveaux de transcription que l'on peut observer *in vivo*. Un « système » supplémentaire est donc indispensable pour la régulation de l'expression des gènes dans les cellules vivantes. En effet, **la régulation de l'expression des gènes requiert l'action concertée de beaucoup de protéines qui se lient à l'ADN et interagissent avec la machinerie basale de transcription, ce qui permet l'activation ou la répression de la transcription en ARNm.** Chez les eucaryotes, les gènes s'expriment très souvent par l'intermédiaire de complexes multiprotéiques composés de divers polypeptides qui se lient à l'ADN de manière séquence-spécifique.

III.1 - Influence de la structure chromatinienne : informations épigénétiques

Dans sa définition moderne, le terme épigénétique désigne des paramètres, héritables au cours des divisions cellulaires, qui contribuent à la régulation d'états fonctionnels au sein d'une cellule sans affecter directement la séquence d'ADN [316]. La chromatine, de par sa structure très condensée, représente une barrière à l'accessibilité de l'ADN par la machinerie basale de transcription et les facteurs de transcription. Elle est en perpétuel équilibre entre état relâché et état condensé suivant les effets antagonistes de multiples complexes protéiques [190] et il ne fait guère de doute que l'état condensé est défavorable à la transcription. L'information épigénétique au sein de la chromatine est principalement véhiculée par des modifications de l'ADN et des histones ; le remodelage de la chromatine constitue donc une étape cruciale dans la régulation de la transcription [408]. Parmi les

[3] Un locus définit l'emplacement d'un allèle ou d'un gène sur un chromosome ou la carte factorielle le représentant. Mais ce terme est généralement employé pour définir l'emplacement précis d'un gène sur un chromosome.

complexes de remodelage de la chromatine, on distingue ceux qui sont susceptibles de modifier les histones elles-mêmes, ceux qui influent sur la méthylation de l'ADN et enfin ceux qui utilisent l'énergie de l'ATP pour modifier la structure du nucléosome [214, 364] (Figure 7).

a. Modifications post-traductionnelles des histones : le « code histone »

Les histones H2A, H2B, H3 et H4 qui forment le cœur du nucléosome sont de petites protéines basiques très conservées au cours de l'évolution. La région la plus conservée de ces histones est leur domaine central structuré qui comprend trois hélices séparées par deux boucles. En revanche, les extrémités N-terminales, et dans une moindre mesure C-terminales, de ces histones sont plus variables, dépourvues de structure secondaire et émergent à la surface du nucléosome (Figure 13a). Ces extrémités sont particulièrement riches en résidus lysine et arginine et donc très basiques. Elles sont la cible de nombreuses modifications post-traductionnelles pouvant affecter leurs charges mais aussi l'accessibilité à l'ADN et les interactions protéines/protéines avec le nucléosome. Les histones sont modifiées par des acétylations, phosphorylations, méthylations et ubiquitinations. Dans la plupart des cas, les sites précis de ces modifications ont été identifiés (Figure 13b) [78, 225, 303].

L'acétylation des histones est la mieux comprise des modifications que peuvent subir ces protéines, aussi bien en terme de résidus affectés qu'en terme de conséquences sur l'activité transcriptionnelle. Plusieurs lysines sur la queue N-terminale de chaque histone peuvent être acétylées par des **histones acétyltransférases (HAT)** et désacétylées par des **histones désacétylases (HDAC)** [226]. Il existe un lien entre l'acétylation des queues des histones et l'activité transcriptionnelle. Les histones hyperacétylées sont associées de manière stable aux régions transcriptionnellement actives et à une structure chromatinienne plus accessible, alors que les histones hypoacétylées se trouvent préférentiellement dans les régions transcriptionnellement silencieuses [99, 162, 163]. En effet, l'acétylation, en détruisant les structures d'ordre supérieur de la chromatine [354], donne à la machinerie basale de transcription et à ses régulateurs un meilleur accès à l'ADN. L'acétylation des queues des histones perturbe la structure du nucléosome en neutralisant les charges positives des lysines, diminuant ainsi son affinité pour l'ADN chargé négativement ou pour les

nucléosomes environnants [152, 241]. Cette acétylation peut aussi influencer la transcription en favorisant ou en empêchant l'interaction avec des facteurs de transcription spécifiques [408].

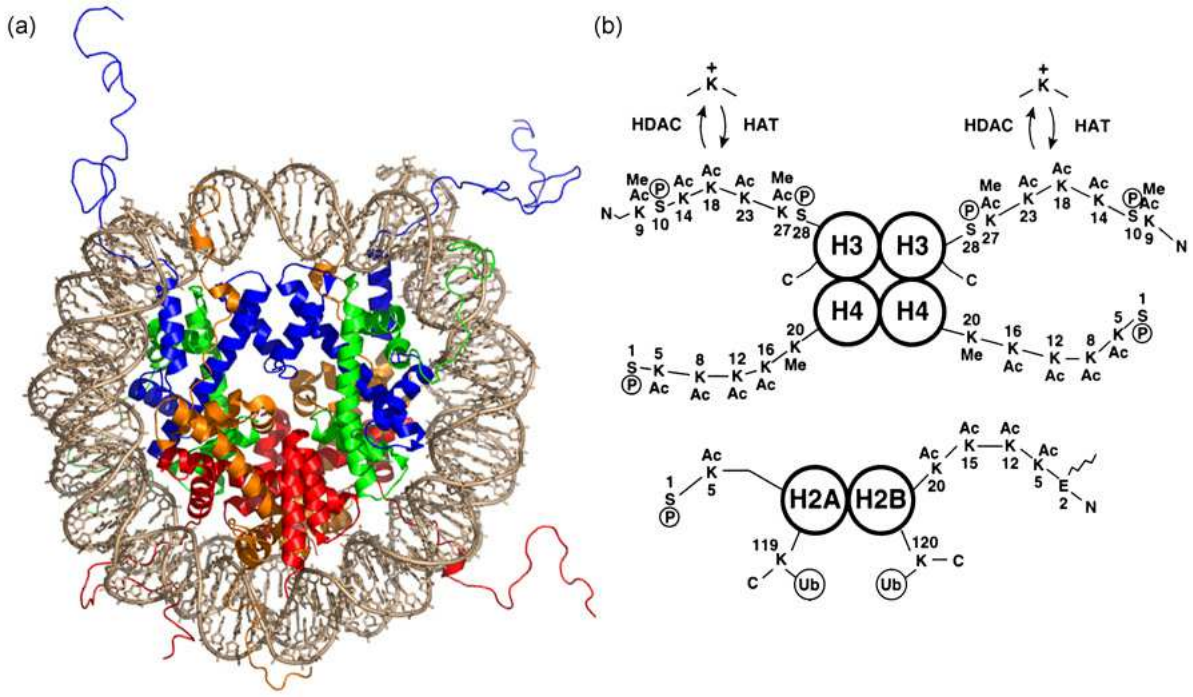


Figure 13. Histones : positionnement au cœur du nucléosome et modifications post-traductionnelles.
(a) Structure du cœur du nucléosome. La double hélice d'ADN (couleur crème) s'enroule autour d'un octamère protéique constitué de deux molécules de chaque histone : H2A (en orange), H2B (en rouge), H3 (en bleu) et H4 (en vert). Cette image a été obtenue à partir de la structure 1KX5 [77] de la base de données PDB. **(b)** Modifications post-traductionnelles des histones. Les histones H3 et H4, montrées sous forme de tétramère, peuvent être acétylées (Ac), méthylées (Me) ou phosphorylées (P). Les histones H2A et H2B, montrées sous forme de dimère, peuvent être modifiées par acétylation, phosphorylation ou ubiquitination (Ub). Les enzymes catalysant une acétylation réversible sont indiquées (histone acétyltransférase : HAT ; histone désacétylase : HDAC). Tirée de Davie *et al.* [78] et adaptée par Marks *et al.* [249].

Les queues des histones peuvent aussi être phosphorylées. La phosphorylation des histones H1 et H3 a été impliquée dans la condensation des chromosomes lors de la mitose [215] et celle de H3 a aussi été liée à une activité transcriptionnelle augmentée [237].

L'ubiquitination des histones H2A et H2B [79] est réversible. Ces histones modifiées sont associées avec un ADN transcriptionnellement actif. L'ubiquitination dépend de la transcription en cours. L'addition d'ubiquitine peut aussi servir à perturber la structure du nucléosome [354] mais le rôle régulateur de cette modification dans la transcription n'est pas encore très bien établi.

Les histones H2B, H3 et H4 peuvent être méthylées mais les effets de cette méthylation sur la transcription sont encore très mal compris. La méthylation est associée avec les formes acétylées de H3 et H4, suggérant ainsi que la méthylation, l'acétylation et l'activation transcriptionnelle sont corrélées.

Les modifications des histones modulent ainsi la structure de la chromatine, permettant de contrôler les fonctions cellulaires liées à l'ADN comme la transcription. Il est proposé que l'ensemble des modifications des histones constituerait un code, appelé « code histone », qui permettrait d'associer à chaque combinaison de modifications un état particulier de la chromatine [182, 382, 383]. La transduction du code histone implique donc la transformation de l'information codée par les modifications d'histones en une réponse contrôlant les fonctions cellulaires liées à l'ADN. Cette transduction peut être la conséquence directe d'un changement de structure de la chromatine, mais peut faire intervenir un intermédiaire protéique, un transducteur, qui interagit avec l'histone modifiée [81, 357]. De plus, l'existence de plusieurs modifications sur une même histone apporte une dimension supplémentaire au code histone. En effet, une modification sur un site peut influencer la capacité d'un autre site à être modifié, de manière synergique ou au contraire antagoniste [183, 246]. Le schéma actuel s'appuie donc sur une combinaison de modifications, et non une modification isolée, afin de conditionner un état particulier de la chromatine. Ce constat met en évidence le caractère très élaboré du code histone, dont le décryptage n'en est qu'à ses prémices.

De plus, il est intéressant de garder en mémoire que chaque histone du cœur du nucléosome, sauf H4, existe dans la cellule de mammifère sous plusieurs formes protéiques, qui ont des similitudes de séquences variables et sont codées par des gènes différents [126]. L'ensemble de ces formes pour chaque histone est regroupé sous le terme de « variants d'histones ». Bien qu'ils soient connus depuis plusieurs dizaines d'années, les variants d'histones n'ont que très récemment fait l'objet d'un regain d'intérêt, en partie grâce aux progrès technologiques.

b. Méthylation de l'ADN

Chez les eucaryotes, seules les cytosines précédant une guanine (dinucléotide CpG) peuvent être méthylées. Les îlots CpG sont définis comme étant des régions de plus de 200 paires de bases, dont le pourcentage en G+C est supérieur à 50% et dont le rapport « fréquence en CpG observée / fréquence en CpG estimée » est supérieur à 0,6 [131]. Leur distribution le long du génome n'est pas homogène : leur fréquence correspond à 1/5ème seulement de la fréquence attendue. Ceci peut être expliqué par le fait que la plupart des dinucléotides CpG de l'ADN sont mutés : la cytosine du dinucléotide CpG est méthylée sur le carbone en position 5 (Figure 6b) par une ADN méthyltransférase. Les îlots CpG correspondent donc à des dinucléotides CpG qui n'ont pas été mutés.

Dans le génome humain, 70 à 80% des dinucléotides CpG sont méthylés mais pas de manière aléatoire : dans les régions où la densité en dinucléotides CpG est faible, ceux-ci sont méthylés alors que dans les régions où la densité en dinucléotides CpG est forte, c'est-à-dire dans les îlots CpG, ceux-ci ne sont pas méthylés. La méthylation des cytosines au niveau des îlots CpG empêche la transcription du gène situé en aval de ces îlots. Des protéines, se liant à l'ADN méthylé, présentent une activité histone désacétylase, ce qui suggère ainsi que ces protéines peuvent convertir la chromatine en état inactif au niveau du site d'initiation de la transcription et donc assurer la quiescence des gènes [187, 273]. Néanmoins, il semblerait que cette méthylation ne soit pas un phénomène réversible : elle permettrait notamment d'inactiver un des deux chromosomes X chez la femme, de réprimer des gènes étrangers comme des gènes viraux, d'inactiver certains gènes lors de la différenciation [186].

c. Remodelage ATP-dépendant

Les nucléosomes sont sujets à un remodelage conformationnel [104] en plus des modifications covalentes que sont les acétylations, phosphorylations, etc. Le remodelage implique la cassure et la reformation des contacts entre les histones et l'ADN. Bien que le mécanisme précis d'un tel remodelage de la chromatine soit encore inconnu, plusieurs complexes de remodelage ont été identifiés dans la plupart des cellules eucaryotes, les plus étudiés étant les complexes SWI/SNF [408] et RSC de la levure et les complexes NURF,

CHRAC et ACF [4] de la drosophile. Tous ces complexes contiennent une sous-unité ATPase avec un domaine hélicase, qui est essentielle pour l'activité de remodelage ainsi que des sous-unités affectant la régulation, l'efficacité et la spécificité de remodelage.

Les différents complexes de remodelage de la chromatine ont la capacité de rendre le nucléosome plus fluide, dans sa position et sa conformation [204], c'est-à-dire de rendre l'ADN plus accessible en créant une structure nucléosomique qui oscille entre la forme d'origine et une nouvelle forme, dite altérée. Lors de ce processus, les histones nucléosomiques peuvent être transférées sur de l'ADN libre. L'état « altéré » rendrait l'ADN plus accessible à des protéines telles que des facteurs de transcription [336]. Néanmoins, l'action des complexes de remodelage ne spécifie pas si l'état de la chromatine obtenu est positif ou négatif pour la transcription. En effet, certains facteurs de remodelage, comme le facteur Swi2, ont été montrés comme ayant un rôle positif pour la transcription de certains gènes et un rôle négatif pour d'autres gènes [362].

III.2 - Les séquences *cis*-régulatrices

Il existe des séquences situées en amont du site d'initiation de la transcription qui contribuent à la régulation de nombreux gènes. Ces séquences peuvent appartenir au **promoteur proximal** ou au **promoteur distal**. Selon le type de facteurs de transcription qui se fixent sur ces séquences de manière spécifique, ces séquences, communément nommées **éléments de régulation**, peuvent être activatrices (amplificateurs ou enhancers) ou répressives (silenceurs ou silencers). En effet, sur les amplificateurs se fixent des activateurs transcriptionnels et sur les silenceurs, des répresseurs transcriptionnels.

Le promoteur proximal (Figure 9) comprend de multiples sites de reconnaissance d'un sous-groupe de facteurs de transcription se fixant à l'ADN de manière séquence-spécifique, dont les facteurs **Sp1**, **NF-Y** et **NF-I**, le premier reconnaissant la boîte GC, les deux autres la boîte CCAAT [33].

[4] SWI/SNF : mating type switching/sucrose non-fermenting ; RSC : remodels the structure of chromatin ; NURF : Nucleosome remodeling factor ; CHRAC : Chromatin accessibility complex ; ACF : ATP-dependent chromatin assembly and remodelling factor.

Néanmoins, la transcription par l'ARN polymérase II est souvent régulée par des sites distants situés à des milliers de paires de bases du site d'initiation de la transcription. En effet, le promoteur distal contient de nombreux éléments de régulation qui sont là pour accomplir une fonction spécifique comme l'activation de la transcription d'un gène dans un type cellulaire bien précis ou à un stade particulier du développement de l'organisme concerné. Ces séquences *cis*-régulatrices augmentent ou répriment la transcription d'un gène indépendamment de leur orientation et de leur distance par rapport au site d'initiation de la transcription. **Ainsi, chaque gène comprend, dans son promoteur au sens large, plusieurs éléments de régulation, chacun contribuant de manière cumulative à la régulation spatiale et temporelle de l'expression de ce gène.**

La première séquence activatrice découverte dans les systèmes de transcription eucaryotes est celle du génome du virus simien SV40 au début des années 80 [18] : elle réside dans une séquence d'environ 100 pb logée à une centaine de bases en amont du site précoce d'amorçage de la transcription de SV40. On la nomme « amplificateur SV40 » car elle augmente la transcription des gènes portés par le même plasmide et ce, qu'elle soit insérée dans n'importe quel sens et en quelque endroit d'un plasmide même placée à des milliers de paires de bases du site d'initiation de la transcription. L'amplificateur SV40 se compose de divers éléments, chacun apportant sa contribution à l'activité globale de l'amplificateur et chacun étant un site de fixation pour une protéine [98].

Sitôt l'amplificateur SV40 découvert, on retrouve des amplificateurs dans d'autres gènes viraux [82] et dans l'ADN des cellules eucaryotes, parfois à au moins 50 kb du promoteur proximal qu'ils gouvernent. L'étude de nombreux activateurs présents dans les cellules eucaryotes révèle qu'ils siègent en amont ou en aval du promoteur proximal, dans un intron, parfois même en aval du dernier exon d'un gène [247].

Beaucoup de ces éléments de régulation sont spécifiques d'un type cellulaire ou encore d'un stade de développement. Ils peuvent être regroupés physiquement et former ainsi des **modules de régulation** dont la définition expérimentale est la suivante : **un module est un fragment d'ADN *cis*-régulateur qui est capable, lorsqu'il est lié à un gène rapporteur et transféré dans une cellule appropriée, d'avoir une fonction régulatrice qui est un sous-ensemble de la fonction régulatrice du promoteur complet** [13]. Les modules contiennent de nombreux sites de liaison de facteurs de transcription qui contribuent de différentes

manières à la régulation générale. Il existe aujourd'hui de nombreux exemples d'organisation des séquences *cis*-régulatrices en modules, l'exemple le plus remarquable étant la régulation de l'expression spatiale et temporelle de plusieurs gènes codant des facteurs de transcription chez l'embryon de drosophile [144, 168].

On s'accorde aujourd'hui sur l'idée que la transcription gouvernée par l'ARN polymérase II est soumise à toute une panoplie de régions régulatrices. D'un côté, on connaît des séquences amplificatrices capables de stimuler la transcription à partir d'un promoteur proximal situé à des milliers de paires de bases ; de l'autre, on a des éléments situés dans le promoteur proximal qui perdent leur pouvoir dès qu'on les éloigne de 15 à 20 bases [257]. De nombreuses régions régulatrices actives sur la transcription à des distances comprises entre ces deux extrêmes ont été identifiées.

III.3 - Les facteurs de transcription

Aux diverses séquences *cis*-régulatrices décrites précédemment viennent se fixer des protéines régulatrices [98]. Ces protéines, communément appelées facteurs de transcription, agissent en général en stimulant la transcription des gènes eucaryotes, mais il existe aussi des facteurs qui la répriment. Les facteurs de transcription liés aux promoteurs proximal et distal régissent l'assemblage du complexe d'amorçage ainsi que la fréquence à laquelle l'ARN polymérase II en place dans ce complexe amorce la transcription. Quand, par exemple, on forme *in vitro* un complexe d'amorçage avec un facteur TFIID dépourvu de ses protomères TAF, la transcription effectuée par ce complexe n'est pas stimulée par les activateurs transcriptionnels fixés sur le promoteur proximal ; ceci démontre que la commande de la transcription exige absolument les protomères TAF (Figure 14) [378].

Les facteurs de transcription reconnaissent leurs sites spécifiques grâce à de petits domaines, très justement appelés domaines de liaison à l'ADN, et coopèrent avec d'autres protéines pour stimuler ou réprimer la transcription par l'intermédiaire d'un domaine effecteur ou domaine de *trans*-activation (Figure 14). Dans la plupart des cas, ces domaines sont interchangeables entre plusieurs protéines ce qui leur donne la caractéristique d'être des unités indépendantes [160]. Différents groupes de facteurs de transcription ont été identifiés

sur la base de la séquence de leurs domaines caractéristiques ainsi que sur leur structure tridimensionnelle et leur façon d'interagir avec l'ADN [160, 293, 369, 392].

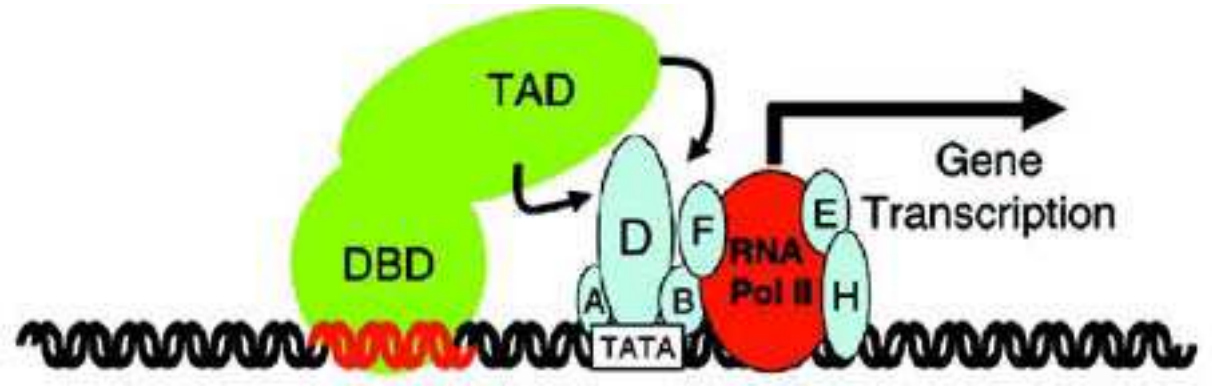


Figure 14. Vue simplifiée de l'interaction entre un facteur de transcription et le complexe de pré-initiation.
L'ARN polymérase II (en rouge) a besoin des facteurs généraux de la transcription (TFII) D, A, B, F, E et H (en bleu) pour reconnaître le site d'initiation de la transcription *via* la boîte TATA (ou d'autres séquences) dans le promoteur proximal. Tous ces facteurs, regroupés sous le terme de complexe de pré-initiation, sont nécessaires pour la transcription basale. Les facteurs de transcription (en vert) se fixent sur des séquences spécifiques situées dans le promoteur grâce à leur domaine de liaison à l'ADN (DBD) et module le niveau de transcription grâce à leur domaine de *trans*-activation (TAD) qui interagit avec le complexe de pré-initiation. Image tirée de l'article de B. Ganss & A. Jheon [130].

Les domaines des facteurs que je vais décrire par la suite, se replient de façon à présenter une protubérance ou une structure flexible, voire les deux, qui entrera en contact avec l'ADN. Ces facteurs montrent des structures très diversifiées permettant le plus souvent l'insertion d'une hélice α dans le grand sillon. Les interactions avec les atomes du squelette sucre-phosphate de l'ADN (liaisons hydrogène et interactions de van der Waals) sont essentielles pour que les domaines se positionnent correctement [160].

Voici une classification des facteurs de transcription selon ce que Stephen C. Harrison a publié en 1991 [160] et selon la base de données dédiée aux facteurs de transcription eucaryotes et à leurs éléments de régulation : TRANSFAC® [209, 356, 404]. Quelques exemples de facteurs dont la structure tridimensionnelle a été déterminée vont être montrés, correspondant en particulier à des facteurs nucléaires recherchés chez *Plasmodium*.

a. Protéines à domaine basique

Le domaine basique de ces protéines, long de 15 à 30 acides aminés, correspond au domaine de liaison à l'ADN proprement dit. Ces protéines appartiennent à différents sous-groupes définis grâce à la structure du domaine qui leur permet de se dimériser. Parmi ces sous-groupes se trouvent notamment les protéines portant une agrafe à leucines (bZIP) [218] ainsi que les protéines à motif hélice-boucle-hélice (bHLH) [267]. Ces protéines ont un rôle important dans la différenciation et le développement. Elles sont aussi intéressantes car elles illustrent parfaitement le rôle important que la formation d'un hétérodimère peut jouer dans la régulation de l'expression des gènes.

Protéines avec une agrafe à leucines (bZIP) (Figure 15a)

Le motif « agrafe à leucines » (leucine zipper en anglais) a été le premier motif conservé dans plusieurs facteurs de transcription eucaryotes à être découvert et aujourd'hui, ce motif apparaît dans une large variété de facteurs de transcription, allant des levures à l'homme [201]. D'une longueur de 30 à 40 résidus, il se situe tout de suite en aval du domaine basique et est caractérisé par une répétition de sept leucines [218].

Les protéines de cette classe peuvent former des homodimères ou des hétérodimères grâce au motif agrafe à leucines et la nature du dimère influe sur l'activité biologique des protéines. Par exemple, le facteur de transcription activateur AP-1 est formé d'une protéine c-Fos et d'une protéine c-Jun [74] ; alors que l'activité du facteur CREB est inhibée lorsqu'il est complexé avec le facteur CREM [122]. Les hétérodimères peuvent aussi acquérir de nouvelles spécificités de liaison à l'ADN et ainsi se fixer à des séquences ADN différentes de celles sur lesquelles se fixent les homodimères [253]. Néanmoins, il apparaît que ces protéines ne forment pas des homodimères très stables et qu'un partenaire convenable est essentiel pour l'hétérodimérisation et la liaison à l'ADN [201].

Protéines à motif hélice-boucle-hélice (bHLH) (Figure 15b)

Les protéines à motif hélice-boucle-hélice [267] ont des similitudes avec la famille présentée précédemment dans le sens où la région basique se fixant à l'ADN est suivie d'une autre région permettant la formation de dimères [387]. C'est cette dernière région qui a

donné son nom à cette classe de protéines car elle est formée d'une hélice α amphipathique [5], d'une boucle puis d'une autre hélice α amphipathique [267].

Tout comme les protéines présentant une agrafe à leucines, les protéines hélice-boucle-hélice ont d'importants rôles dans la différenciation et le développement et leur activité est modulée par la formation d'hétérodimères mélangeant des activateurs, des protéines ubiquitaires et des répresseurs [20].

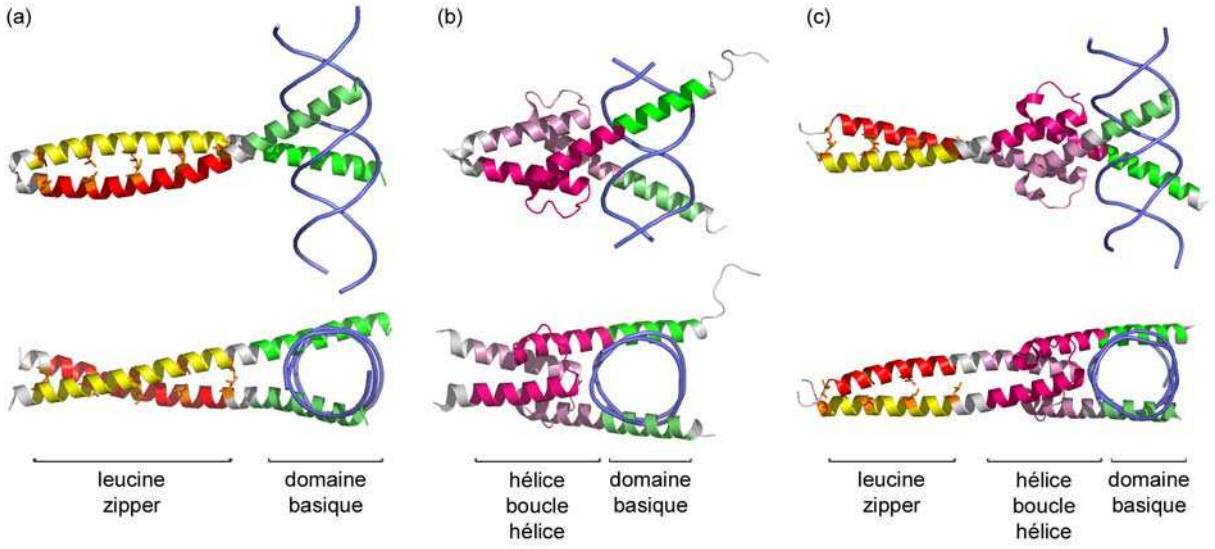


Figure 15. Protéines ayant un domaine basique comme domaine de liaison à l'ADN.
Chaque structure est montrée sous deux angles différents. Tous ces facteurs de transcription sont composés de deux domaines : un domaine de liaison à l'ADN et un domaine qui leur permet de se dimériser. Le domaine basique (en vert) est, pour toutes ces structures, le domaine qui se lie à l'ADN (en indigo). Les protéines se distinguent donc par leur domaine de dimérisation : soit un domaine agrafe à leucines (en rouge et jaune), soit un motif hélice-boucle-hélice (en rose foncé et clair), soit les deux. **(a)** Le facteur AP-1 humain est un hétérodimère formé de la protéine Jun (en jaune et vert) et de la protéine Fos (en rouge et vert). Sur les domaines agrafe à leucines sont matérialisées en orange les chaînes latérales des leucines importantes pour la dimérisation. **(b)** Le facteur MyoD de souris s'homodimérise pour se lier à l'ADN. **(c)** Le facteur de transcription formé des protéines Myc (en jaune, rose foncé et vert) et Max (en rouge, rose clair et vert clair) s'hétérodimérise grâce au motif hélice-boucle-hélice et aux domaines agrafe à leucines. Sur les domaines agrafe à leucines sont matérialisées en orange les chaînes latérales des leucines importantes pour la dimérisation. Les fichiers PDB utilisés pour faire ces images sont les suivants : (a) 1FOS [142], (b) 1MDY [244], (c) 1NKP [269].

[5] Dans une hélice amphipathique, les résidus hydrophiles sont situés sur une face tandis que les résidus hydrophobes sont situés sur l'autre face.

Protéines à motif hélice-boucle-hélice et à agrafe à leucines (bHLH-ZIP) (Figure 15c)

Les protéines à motif hélice-boucle-hélice et agrafe à leucines sont ce que l'on pourrait considérer comme une chimère des deux précédents types de protéines. Le domaine de liaison à l'ADN qui reste une région basique est suivi par un motif hélice-boucle-hélice puis par une agrafe à leucines. La dimérisation se fait à la fois par les domaines en hélice-boucle-hélice et par les domaines en agrafe à leucines. La fonction des hétérodimères formés par des protéines à motif bHLH-ZIP dépend de la nature des protomères : l'hétérodimère formé par les protéines Myc et Max a un potentiel oncogénique fort, tandis que l'hétérodimère Mad-Max agit en tant que répresseur transcriptionnel [175].

b. Protéines à motif en doigt de zinc (Zn finger)

Les motifs en doigt de zinc (Figure 16) sont d'autres motifs structuraux majeurs impliqués dans les interactions ADN-protéine [27, 193]. Les protéines comportant un motif en doigt de zinc sont impliquées dans plusieurs aspects de la régulation des gènes eucaryotes : la différenciation et la croissance, les proto-oncogènes, les facteurs généraux de la transcription, etc. [26, 27, 207]. Au départ, on pensait que toutes les protéines de cette famille présentaient le même schéma, à savoir des répétitions en tandem du motif en doigt de zinc, chacun comprenant la séquence suivante : Tyr/Phe-X-Cys-$X_{2 \text{ ou } 4}$-Cys-X_{12}-His-$X_{3\text{-}5}$-His (X pour n'importe quel acide aminé). Or il s'est avéré que cette séquence représentait une sous-classe des protéines à motif en doigt de zinc, la sous-classe Cys2His2 ou C_2H_2. Aujourd'hui, on connaît d'autres sous-classes comme, par exemple, la sous-classe Cys4 qui regroupe des protéines régulatrices du type récepteur d'hormones ou la sous-classe Cys6 qui rassemble des régulateurs métaboliques.

Le doigt de zinc C_2H_2 est un site de liaison compact de 25-30 résidus. Il est composé d'une boucle β et d'une hélice α qui sont « pelotonnées » l'une contre l'autre grâce à un ion Zn^{2+} en liaison avec deux cystéines de la boucle β et deux histidines de l'hélice α [369]. Bien que souvent associé aux facteurs de transcription se fixant sur le promoteur distal, le motif en doigt de zinc peut aussi se fixer au niveau du promoteur proximal, ce qui est le cas pour les facteurs Sp1 (Figure 16a) [41] et YY1 (Figure 16b) [384, 385].

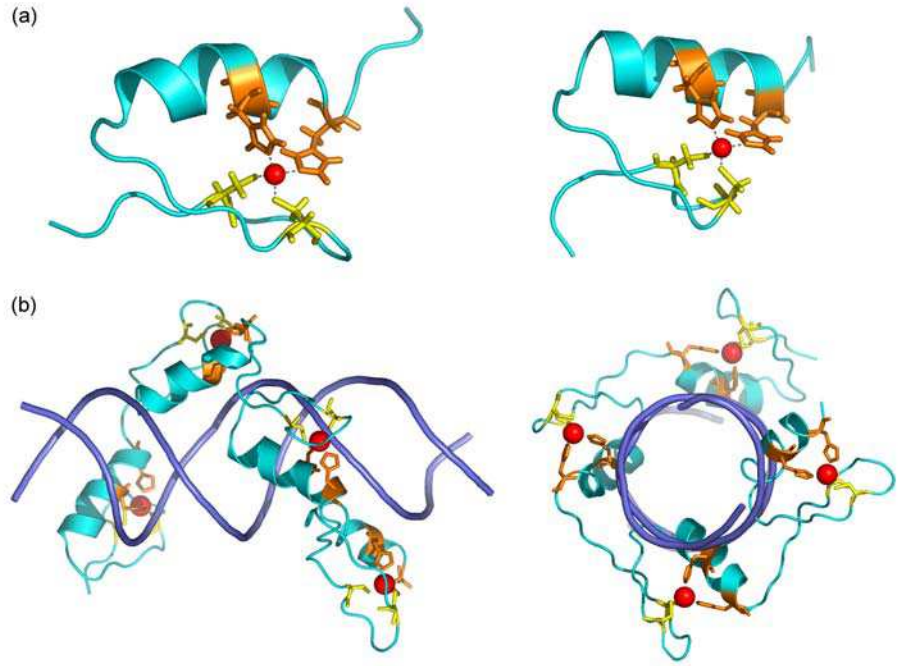


Figure 16. Protéines ayant un motif en doigt de zinc dans leur domaine de liaison à l'ADN.
Le motif en doigt de zinc est composé d'une boucle β lovée contre une hélice α. Un ion Zn^{2+} (en rouge) maintient la structure en interagissant avec les chaînes latérales de deux cystéines de la boucle (en jaune) et de deux histidines de l'hélice (en orange). **(a)** Deuxième et troisième domaines en doigt de zinc de la protéine Sp1 humaine (respectivement à gauche et à droite). Les liaisons entre l'atome de zinc et les chaînes latérales de deux cystéines (en jaune) et de deux histidines (en orange) sont indiquées par des pointillés. **(b)** Le facteur YY1 humain comporte quatre domaines en doigt de zinc qui s'enroulent autour de l'ADN (en indigo) en suivant le grand sillon. Les fichiers PDB utilisés pour faire ces images sont les suivants : (a) 1SP2 et 1SP1 [274], (b) 1UBD [173].

c. Protéines à motif hélice-tour-hélice (HTH)

La structure en motif hélice-tour-hélice a d'abord été découvert chez les procaryotes puis chez les eucaryotes [161], et aujourd'hui de nombreuses structures HTH ont été déterminées que ce soit pour des protéines seules ou des protéines complexées avec un ADN [293]. Dans sa forme la plus simple, le motif HTH a une longueur d'environ 20 résidus et est constitué, comme son nom l'indique, de deux hélices α quasiment perpendiculaires connectées par un tour de 4 résidus [210], la deuxième hélice étant celle qui se loge dans le grand sillon et crée des interactions spécifiques avec l'ADN [160, 161]. Des variations du domaine de liaison HTH classique ont été découvertes ces dernières années au niveau du tour entre les deux hélices : en effet, celui-ci peut être plus long et ainsi adopter une conformation différente. Le nombre de résidus insérés dans le tour est assez variable, allant de un seul chez le proto-oncogène *c-myb* [284] à 21 chez le facteur de transcription hépatocytaire LFB1/HNF1 [116].

Des résidus hydrophobes conservés dans les hélices ainsi qu'une glycine dans le tour participent à la stabilisation de l'arrangement entre les deux hélices et permettent à la structure HTH de rester compacte par rapport au reste de la protéine [293]. Cependant, le motif HTH ne peut se replier ou fonctionner tout seul : ce n'est qu'une partie d'un domaine de liaison à l'ADN plus grand qui peut présenter un environnement structural différent d'une protéine à l'autre [363]. C'est ainsi que les protéines contenant un motif HTH peuvent être classifiées en plusieurs groupes structuraux selon la nature et la position des autres éléments structuraux (hélices α, brins β ou structures en épingle à cheveux) qui forment avec le motif HTH le cœur hydrophobe des protéines [161]. Voici deux de ces groupes.

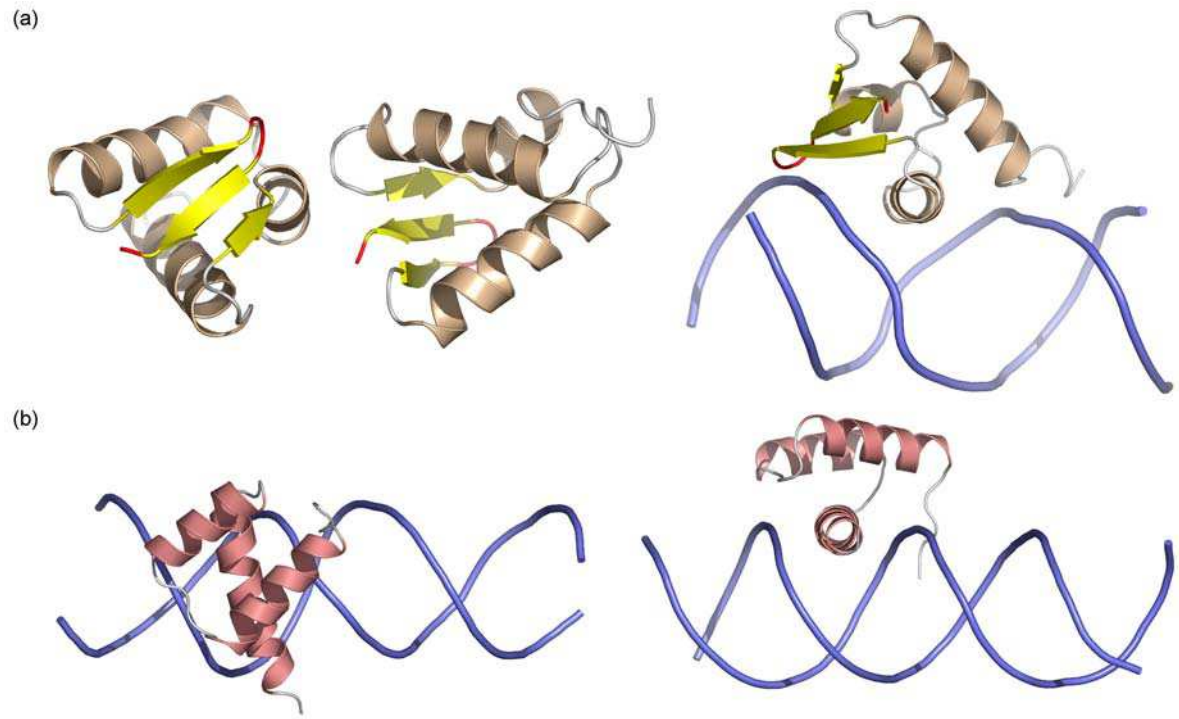


Figure 17. Protéines ayant un domaine de liaison à l'ADN à motif hélice-tour-hélice.
(a) Trois vues du domaine de liaison à l'ADN de la protéine E2F4 humaine. Les différents constituants du motif à hélices ailées sont représentés en couleur ivoire (pour les hélices α), en jaune (pour les brins β) et en rouge (pour les ailes). Pour plus de clarté, la protéine DP-2, qui forme un hétérodimère avec E2F4, n'a pas été représentée. **(b)** Deux vues de l'homéodomaine du facteur Engrailed de la drosophile fixé à une séquence d'ADN (en indigo). Les structures composant l'homéodomaine sont différenciées grâce aux couleurs : saumon pour les hélices α et blanc pour les tours. Les fichiers PDB utilisés pour faire ces images sont les suivants : (a) 1CF7 [417] et (b) 3HDD [124].

Les protéines à « hélices ailées » (Figure 17a)

Le facteur HNF-3 fut le premier facteur comportant des hélices ailées à être découvert en 1993 [61]. Au niveau topologique, le motif à hélices ailées est une structure α/β compacte

composée de deux ailes (A_1 et A_2), 3 hélices α (H_1, H_2 et H_3) et trois brins β (B_1, B_2 et B_3) arrangés de la manière suivante : H_1-B_1-H_2-H_3-B_2-A_1-B_3-A_2 [129]. Les protéines à hélices ailées diffèrent du modèle canonique des protéines HTH par la longueur du « tour » situé entre les hélices H_2 et H_3.

Comme les autres facteurs de transcription, ils peuvent agir seul ou sous forme de dimères. En effet, on retrouve ce motif dans les facteurs de transcription de la famille E2F qui contrôle les gènes impliqués dans la croissance et la réplication de l'ADN [346]. La liaison à l'ADN des protéines E2F est augmentée quand celles-ci se dimérisent avec des membres de la famille DP, eux-mêmes étant des facteurs à hélices ailées [417].

Les protéines présentant un homéodomaine (Figure 17b)

L'homéodomaine est un motif de liaison à l'ADN présent dans une grande famille de régulateurs eucaryotes [137, 340]. Même si les séquences conservées ont été identifiées au début dans des protéines régulant le développement chez la drosophile, aujourd'hui on sait que l'homéodomaine a un rôle plus général dans la régulation des gènes chez les eucaryotes.

En clonant et en séquençant des gènes homéotiques de la drosophile très bien connus (*antennapedia*, *ultrabithorax*, *engrailed*, *fushi tarazu*), on remarqua que les protéines comportaient la même région de 60 acides aminés bien conservés. Contrairement à une structure HTH seule, ces 60 résidus, qui prirent le nom d'homéoboîte ou homéodomaine, forment une structure repliée stable, composée de trois hélices α – les deux premières sont empaquetées l'une contre l'autre de manière anti-parallèle alors que la troisième est à peu près perpendiculaire aux deux autres – et surtout capable de se fixer à l'ADN par elle-même [2].

d. Les protéines à architecture β en contact avec le petit sillon

Il est difficile de trouver une caractéristique commune aux domaines de liaison à l'ADN de cette classe de facteurs de transcription. Par exemple, les domaines de liaison à l'ADN des facteurs de transcription HMGB ne contiennent que des hélices alors que ceux des protéines composant le facteur NF-κB (Figure 18a) ou encore ceux de la protéine p53 (Figure 18b) ou de la TBP (Figure 18c) contiennent des hélices α et des brins β. Les protéines HMGB et la TBP

se retrouvent ensemble dans cette classe à cause leur mode d'interaction avec l'ADN : une insertion dans le petit sillon qui entraîne une vrille prononcée de la double hélice [356].

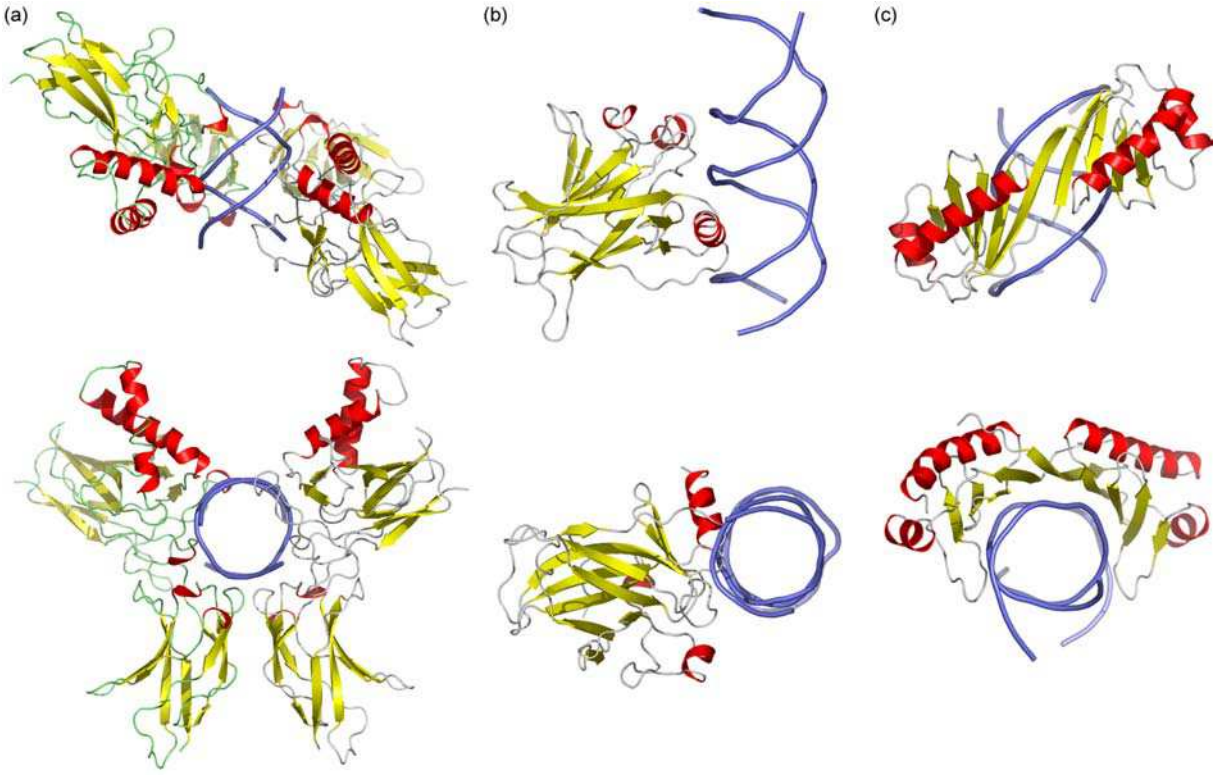


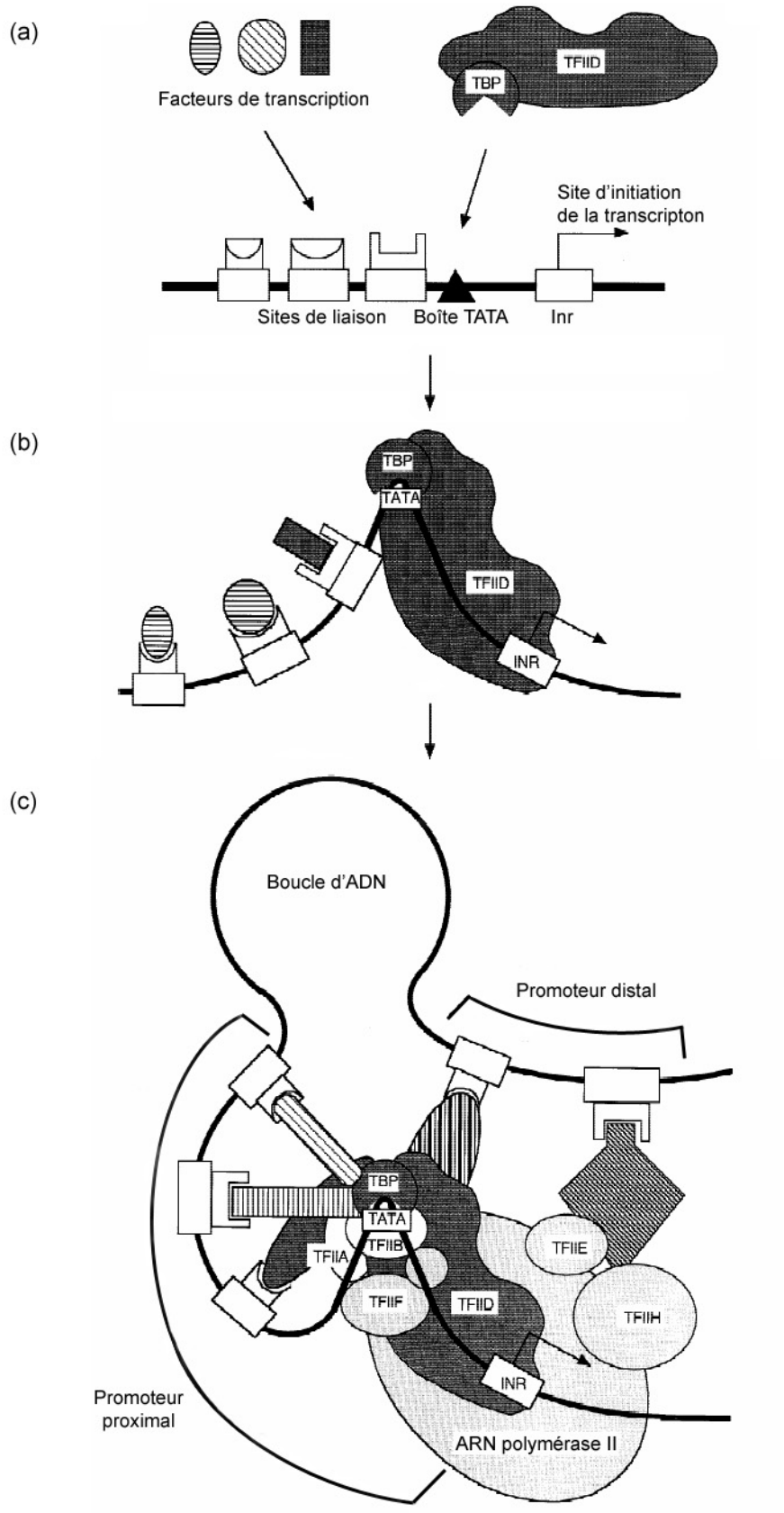
Figure 18. Protéines à architecture β en contact avec le petit sillon.
Les structures sont représentées selon le même mode : les hélices α en rouge, les brins β en jaune et l'ADN en indigo. **(a)** Homodimère de la protéine NF-κB p52 humaine. **(b)** Suppresseur de tumeur p53. **(c)** TATA-binding protein. Les fichiers PDB utilisés pour faire ces images sont les suivants : (a) 1A3Q [71], (b) 1TUP [59] et (c) 1CDW [279].

III.4 - Modèle de mécanisme d'activation de la transcription spécifique d'un gène

Les enzymes modifiant la chromatine, les facteurs généraux de la transcription ainsi que les facteurs de transcription, qu'ils soient activateurs ou répresseurs, s'unissent pour qu'un gène soit transcrit au niveau auquel il doit être transcrit pour le bon fonctionnement de la cellule. L'activation de la transcription d'un gène dépend donc de l'assemblage de complexes multiprotéiques stéréospécifiques sur les différentes parties des promoteurs [379]. Des analyses par cristallographie aux rayons X de complexes comprenant des facteurs de transcription liés à de l'ADN se sont concentrées exclusivement sur les cas où les sites de liaison des facteurs de transcription interagissant ensemble et agissant de manière

synergique sont très proches les uns des autres sur un promoteur [405]. Néanmoins dans beaucoup de gènes eucaryotes, les facteurs de transcription se fixent sur le promoteur à une certaine distance et interagissent quand même ensemble pour activer la transcription [54]. Les boucles formées par l'ADN permettent de rapprocher les facteurs de transcription éparpillés le long de l'ADN suffisamment pour qu'ils puissent former un complexe multiprotéique [334, 335] (Figure 19).

Figure 19. Assemblage du complexe multiprotéique contrôlant la transcription (→).
(a) Représentation schématique d'un promoteur proximal composé des sites de liaison de trois facteurs de transcription, de la boîte TATA et de la région Inr. **(b)** Une fois la chromatine restructurée de façon à rendre l'ADN accessible, le facteur TFIID, composé de la TATA-binding protein (TBP) et de ses co-facteurs se fixe à la boîte TATA. La fixation de la TBP sur la boîte TATA entraîne une courbure de l'ADN d'environ 90°. Les facteurs de transcription se fixent sur les séquences *cis*-régulatrices qui leur sont propres : les activateurs se fixent sur des amplificateurs et augmentent le taux de transcription tandis que les répresseurs se fixent sur des silenceurs et interfèrent avec l'action des activateurs en réduisant le taux de transcription. **(c)** Les autres facteurs généraux de la transcription (TFII A/B/F/E/H) ainsi que l'ARN polymérase II rejoignent ensuite le promoteur proximal. L'ADN forme alors des boucles pour que les facteurs de transcription du promoteur proximal comme du promoteur distal, qu'ils soient activateurs ou répresseurs, interagissent avec les co-facteurs de la TBP. Ces co-facteurs transmettent alors les informations reçues des facteurs de transcription aux facteurs généraux de la transcription qui, en réponse, positionnent l'ARN polymérase II au début de la séquence codante et la mettent en route. Adapté à partir d'un article de Thomas Werner [401].


(a)
Facteurs de transcription
TFIID
TBP
Site d'initiation
de la transcripton
Sites de liaison
Boîte TATA
Inr
(b)
TBP
TATA
TFIID
INR
(c)
Boucle d'ADN
Promoteur distal
TBP
TATA
TFIIB
TFIIA
TFIIE
TFIIF
TFIID
TFIIH
INR
Promoteur
proximal
ARN polymérase II

ETAT DE L'ART SUR *PLASMODIUM FALCIPARUM*

Lorsque j'ai commencé ce travail, très peu de choses étaient connues sur les éléments pouvant intervenir dans la transcription et la régulation de la transcription chez le parasite *P. falciparum*.

L'expression des gènes de Plasmodium falciparum

Les gènes de *P. falciparum* sont monocistroniques [220] et contiennent des introns dans les séquences codantes ; les sites donneurs et accepteurs d'épissage sont conformes à ce que l'on rencontre chez les autres eucaryotes [333, 397]. Les phases ouvertes de lecture sont flanquées en amont d'une région 5' non traduite et en aval d'une région 3' non traduite contenant un ou plusieurs sites de polyadénylation [220]. De plus, les ARNm sont coiffés en plus d'être polyadénylés [220, 355, 402]. Tout ceci est donc conforme à ce que l'on connaît de la structure des gènes chez les eucaryotes (Figure 8).

Tous les gènes du parasite sont exprimés de manière différente : certains sont exprimés de manière constitutive tandis que d'autres sont exprimés lors d'un (de) certain(s) stade(s) de son développement. Lorsque je suis arrivée au laboratoire, cette spécificité de stade avait été observée chez plusieurs antigènes de surface comme les protéines :

- CSP (circumsporozoite protein) [103],
- GBP130 (glycophorin-binding protein) [301, 314],
- HRPII (histidine-rich protein) [400],
- KAHRP (knob-associated histidine-rich protein) [307],
- RESA (ring-infected erythrocyte surface antigen) [42],
- MSP-1 (merozoite surface protein) [157].

On pouvait s'attendre à ce que l'expression de certains gènes soit régulée en fonction du développement : par exemple, le produit du gène *msp-1*, comme son nom l'indique, est une protéine de surface requise pour la formation des mérozoïtes. Mais cette spécificité d'expression a aussi été observée pour des gènes de ménage et même certains ARNr. En effet, *Plasmodium falciparum* possède deux gènes codant l'actine : l'un est exprimé de manière constitutive tandis que l'autre n'est exprimé que lors des stades sexués [402]. De la même manière, il existe deux gènes codant la tubuline α : le premier est exprimé tout au long du développement alors qu'on ne trouve le deuxième que dans les gamétocytes mâles [88, 315]. Quant aux ARNr, ils ont été classés en différents types [230, 393, 394] : les ARNr de type A

sont exprimés lors des stades hépatiques tardifs, de la schizogonie érythrocytaire et lors de la différenciation sexuée tandis que les ARNr de type S sont exprimés lors de la sporogonie et lors des stades hépatiques précoces.

Des études ont aussi montré que des gènes, qui sont fonctionnellement liés, peuvent être régulés de la même manière lors du développement parasitaire. C'est ainsi le cas de gènes codant des éléments de la machinerie de réplication de l'ADN : les gènes codant la grande sous-unité de l'ADN polymérase δ [171], la protéine PCNA (proliferating cell nuclear antigen, [171]) et les topo-isomérases I et II [58] sont tous exprimés dans les trophozoïtes tardifs et les schizontes, leur pic d'expression correspondant avec le début de la réplication de l'ADN [403].

Les éléments intervenant dans la régulation de l'expression des gènes

Alors que seule l'histone H2A avait été identifiée chez *P. falciparum* en 1992 [72], des expériences ont montré, dès 1994, que l'organisation de base de la chromatine était, chez *Plasmodium*, similaire à ce qui avait été observé chez d'autres eucaryotes [55], c'est-à-dire une organisation en nucléosomes. Ont ensuite été identifiées les histones H2B et H3 en 1995 [24, 240] et enfin l'histone H4 en 1997 [23, 239]. Et parmi les protéines modifiant les histones (voir p. 56), une HDAC a été caractérisée dans le génome en 1999 [188].

L'ARN polymérase II est une enzyme constituée de plusieurs sous-unités. La plus grande sous-unité interagit avec l'ADN et contient le domaine C-terminal qui sera phosphorylé par le facteur TFIIH, l'énergie libérée par cette phosphorylation étant utilisée pour séparer les deux brins d'ADN (voir p. 50). Chez *P. falciparum*, le gène codant cette grande sous-unité de l'ARN polymérase II a été identifié sur le chromosome 3 [231]. De la même manière, les gènes codant les grandes sous-unités des ARN polymérases I [123] et III [232] ont été caractérisés sur les chromosomes 9 et 13.

Parmi tous les facteurs généraux de la transcription qui sont associés à l'ARN polymérase II, seule la TBP a été identifiée chez *P. falciparum*. Deux gènes codant la TBP ont été annotés dans le génome du parasite, sur les chromosomes 5 [256] et 12 [165] ; tous deux sont transcrits lors de la phase érythrocytaire du développement parasitaire, mais celui du chromosome 12 l'est plus particulièrement dans le trophozoïte tardif [165].

Il existe de nombreuses preuves indiquant que les promoteurs de *P. falciparum* présentent, comme chez les autres eucaryotes, une structure bipartite, à savoir un promoteur proximal responsable de l'initiation de la transcription [70], contrôlée en amont par des éléments de régulation, situés dans les promoteurs proximal et distal, qui diffèrent d'un promoteur à l'autre en composition et en localisation par rapport au site d'initiation de la transcription (Figure 20). En effet, il a été montré que les gènes étaient monocistroniques ce qui implique la présence de régions régulatrices autour des séquences codantes permettant à la transcription de débuter et de se terminer [4, 220]. Ensuite, les séquences situées en amont des phases ouvertes de lecture de *Plasmodium* semblent comporter des motifs homologues à des sites de liaison de facteurs de transcription eucaryotes communs, même si leur fonction dans l'activité du promoteur n'a pas été établie [171, 220, 221, 360]. Enfin, des retardements sur gel ont montré que des protéines nucléaires se fixaient sur de petites séquences d'ADN de manière spécifique et, parfois, en fonction du stade parasitaire utilisé pour faire les extraits nucléaires [219, 220]. Cependant, au moment où j'ai commencé ce travail, aucune de ces protéines nucléaires n'avait été identifiée ce qui fait **qu'aucun facteur de transcription n'était connu chez *Plasmodium falciparum***.

Toutes ces données suggèrent que les régions situées en amont des gènes plasmodiaux contiennent au moins le minimum nécessaire à l'initiation de la transcription. Cependant, la richesse en A+T de ces régions rend très difficile la localisation d'une boîte TATA.

Il est intéressant de souligner que les deux TBP annotées chez *P. falciparum* divergent assez des protéines des autres eucaryotes, ceci pouvant être une réponse à la composition particulière du génome et notamment des régions intergéniques. On peut donc penser qu'à cause de cette richesse en A+T, le parasite aurait développé une série de facteurs de transcription uniques, c'est-à-dire distincts de ce que l'on trouve chez la levure ou les eucaryotes supérieurs, en plus de facteurs de transcription communs à tous les eucaryotes. Ceci expliquerait le fait que des séquences *cis*-régulatrices sans aucune homologie avec des sites de liaison connus chez les autres eucaryotes aient été identifiées dans les promoteurs de certains gènes parasitaires comme ceux présentés dans la Figure 20.

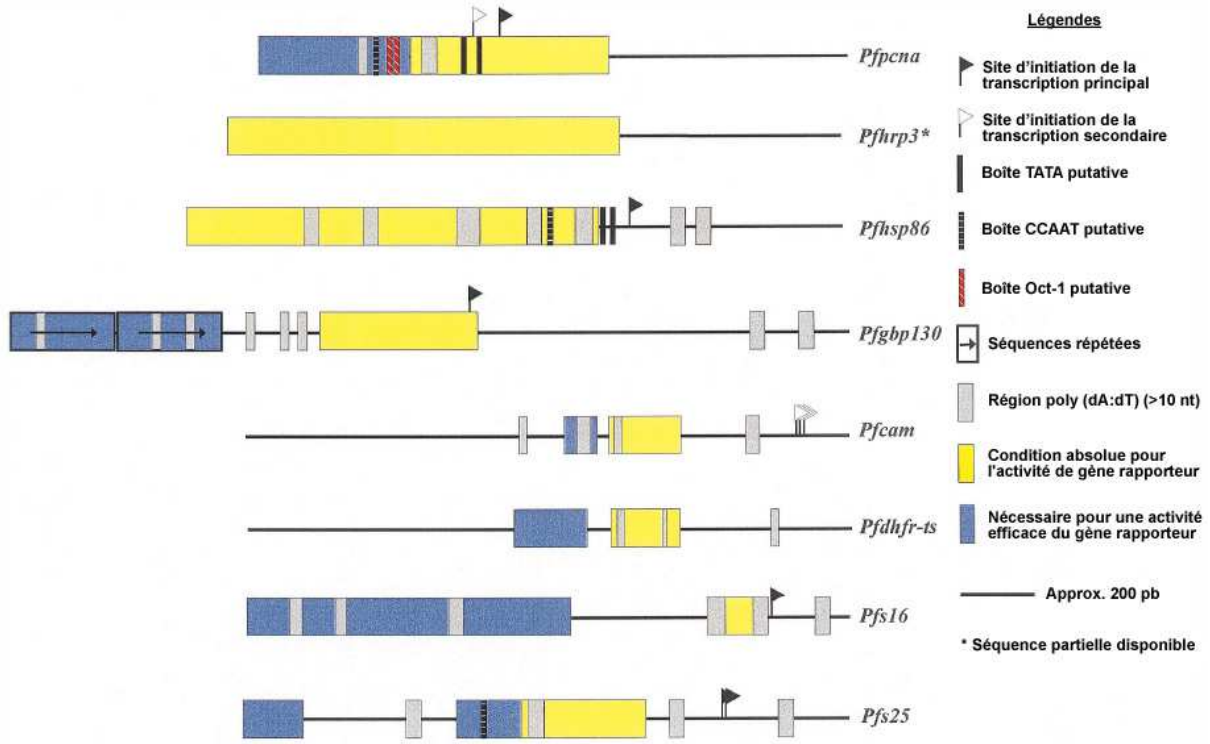


Figure 20. Schémas des régions promotrices de quelques gènes de *Plasmodium falciparum*.
Abréviations : proliferating cell nuclear antigen (*Pfpcna*, [172]) ; histidine-rich protein 3 (*Pfhrp3*, [410]) ; heat shock protein 86 (*Pfhsp86*, [360, 410]) ; glycophorin-binding protein 130 (*Pfgbp130*, [220]) ; calmodulin (*Pfcam*, [70, 324]) ; dihydrofolate reductase-thymidine synthetase (*Pfdhfr-ts*, [70]) ; sexual stage antigen (*Pfs16*, [85]) ; sexual stage antigen 25 (*Pfs25*, [83]). D'après P. Horrocks *et al*. [170].

Comme le génome de *Plasmodium falciparum* présente des caractéristiques comparables à celles des génomes des autres eucaryotes, il était logique de penser que les gènes du parasite répondaient au mécanisme transcriptionnel général des eucaryotes. De plus, le parasite passe d'un hôte vertébré à un hôte invertébré (Figure 3) et révèle des étapes de développement qui sont morphologiquement très distinctes. Ceci nous indique donc que **l'expression des gènes doit être fortement régulée dans les parasites**. Les différentes étapes du développement du parasite impliquent une **régulation coordonnée des gènes** qui se fait à plusieurs niveaux : épigénétique, transcriptionnel, post-transcriptionnel, traductionnel et post-traductionnel. Ainsi, une étape du développement est caractérisée par **l'expression d'une palette de gènes**, distincte de celle qui est observée à une autre étape du cycle parasitaire.

L'objectif de cette thèse a donc été d'identifier les deux partenaires de la régulation transcriptionnelle chez *Plasmodium falciparum* dans le but de comprendre la première étape de la régulation des gènes lors de la phase érythrocytaire du développement du parasite. L'interaction de ces deux partenaires ADN-protéines de la régulation transcriptionnelle est

cruciale pour la survie des parasites. D'un côté, les éléments agissant en *cis*, c'est-à-dire les **éléments de régulation** qui se situent en amont des séquences codantes ; de l'autre, les éléments agissant en *trans*, c'est-à-dire les **facteurs de transcription**, qui se fixent chacun sur un élément de régulation qui lui est spécifique, ainsi que des facteurs nucléaires spécifiques de structures d'ADN. **La présence, la fréquence et la position de certains éléments de régulation en amont d'un gène, ainsi que la présence ou l'absence de facteurs de transcription à un moment donné du développement font qu'un gène est exprimé ou non.** Or lorsque le parasite se développe chez l'homme, notamment dans ses globules rouges, stade responsable des fièvres chroniques de la maladie, on se rend compte qu'il suit un programme très défini et que chaque gène doit être exprimé à un moment précis du développement.

MATERIELS ET METHODES

I - Les différents outils utilisés

Bases de données

EMBL	http://www.ebi.ac.uk/embl/
GenBank®	http://www.ncbi.nlm.nih.gov/Genbank/index.html
TRANSFAC®	http://www.gene-regulation.com
SwissProt, TrEMBL	http://www.expasy.org/sprot/
PIR	http://pir.georgetown.edu/
UniProt	http://www.expasy.uniprot.org/
PROSITE	http://www.expasy.org/prosite/
Pfam	http://www.sanger.ac.uk/Software/Pfam/
PlasmoDB	http://plasmodb.org
dictyBase	http://dictybase.org/
Medline / PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
OMIM	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
iHOP	http://www.pdg.cnb.uam.es/UniPub/iHOP/ www.ihop-net.org/UniPub/iHOP/
PDB	http://pdbprod2.sdsc.edu/pdb/Welcome.do
PDBSum	http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/
CATH	http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html
SRS	http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+top+-newId

Prédiction du site d'initiation de la transcription

NNPP 2.2	http://www.fruitfly.org/seq_tools/promoter.html
Promoter 2.0	http://www.cbs.dtu.dk/services/Promoter/
PromoterScan	http://thr.cit.nih.gov/molbio/proscan/
TSSG, TSSW, TSSP	http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=promoter

Recherche de motifs dans les séquences nucléiques

PATTERNn	http://www.infobiogen.fr/services/analyseq/cgi-bin/patternn_in.pl
AlignACE	http://atlas.med.harvard.edu/
MEME	http://meme.sdsc.edu/meme/meme.html
GIBBS Motif Sampler	http://bayesweb.wadsworth.org/cgi-bin/gibbs.7.pl?data_type=DNA
MatInspector	http://www.genomatix.de/

Visualisation des séquences d'éléments de régulation et des promoteurs

WebLogo	http://weblogo.berkeley.edu/
Feature map	http://rsat.ulb.ac.be/rsat/

Comparaison de séquences

BLAST2	http://www.infobiogen.fr/services/analyseq/cgi-bin/blast2_in.pl
	http://www.ebi.ac.uk/blastall/index.html
	http://plasmodb.org/plasmodb/servlet/sv?page=blast
PSI-BLAST	http://www.infobiogen.fr/services/analyseq/cgi-bin/pp-blast_in.pl
LFASTAp, LALIGNp	http://www.infobiogen.fr/services/analyseq/cgi-bin/lfastap_in.pl
LFASTAn, LALIGNn	http://www.infobiogen.fr/services/analyseq/cgi-bin/lfastan_in.pl

Alignement de séquences

DiAlign2	http://www.infobiogen.fr/services/analyseq/cgi-bin/dialign2_in.pl
ClustalW	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_clustalw.html
MultAlin	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_multalin.html
	http://prodes.toulouse.inra.fr/multalin/multalin.html

Annotation des protéines et identification de domaines

PSORT	http://psort.nibb.ac.jp/form.html
MotifScan	http://myhits.isb-sib.ch/cgi-bin/motif_scan
NetPhos2.0	http://www.cbs.dtu.dk/services/NetPhos/

Phylogénie

Package PHYLIP	http://evolution.genetics.washington.edu/phylip/getme.html

Prédiction de structures secondaires

Consensus	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html

Prédiction de structures tertiaires : modélisation par homologie

Meta-Server @TOME	http://bioserv.cbs.cnrs.fr/HTML_BIO/frame_meta.html
TITO	http://bioserv.cbs.cnrs.fr/HTML_BIO/frame_tito.html
Modeller	http://salilab.org/modeller/

Qualité et comparaison des structures tridimensionnelles

Verify3D	http://www.doe-mbi.ucla.edu/Services/Verify_3D/
ProSa2003	http://www.came.sbg.ac.at/Services/prosa.html
CE	http://cl.sdsc.edu/ce/ce_align.html
ProFit	http://www.bioinf.org.uk/software/profit/
SCit	http://bioserv.rpbs.jussieu.fr/cgi-bin/SCit
SCWRL3.0	http://dunbrack.fccc.edu/SCWRL3.php

Visualisation des structures tridimensionnelles

SwissPDB Viewer	http://www.expasy.org/spdbv/
PyMOL	http://pymol.sourceforge.net/

II - Les matériels et méthodes utilisées

II.1 - Sélection des régions intergéniques

Lorsque le génome nucléaire de la souche 3D7 de *P. falciparum* n'était pas encore complètement séquencé, la recherche des régions intergéniques commençait par l'identification de séquences codantes dans les parties déjà séquencées du génome (chromosomes en cours d'assemblage, contigs et shotguns non encore assemblés). Ceci a permis de savoir sur quel chromosome se situait chaque gène et de sélectionner les 5 000 nucléotides situés en amont de l'ATG initiateur quand cela était possible, c'est-à-dire quand le gène ne se trouvait pas à une des extrémités du contig ou du shotgun. Sur ce fragment d'ADN, des phases ouvertes de lecture potentielles ont été recherchées ; ainsi, des séquences intergéniques putatives situées entre l'ATG initiateur du gène d'intérêt et l'ATG ou le codon-stop (selon l'orientation) du gène situé en amont ont pu être identifiées (Figure 21).

Une fois que le génome a été entièrement séquencé [132], le travail a été simplifié à l'extrême grâce à la mise en place d'une base de données dédiée au génome du parasite : PlasmoDB [205] et à tous les outils permettant de manipuler les séquences.

Comme le site d'initiation de la transcription n'est pas connu pour la plupart des gènes étudiés et que les programmes permettant de le situer précisément qui existent sur la toile ne sont pas optimisés pour un génome aussi particulier que celui de *P. falciparum*, la région 5'UTR du gène d'intérêt ainsi que la région 5' ou 3' UTR du gène situé en amont ont été conservées. De plus ces régions, notamment la région 5'UTR du gène d'intérêt, pourraient contenir des séquences impliquées dans la régulation de la transcription du gène.

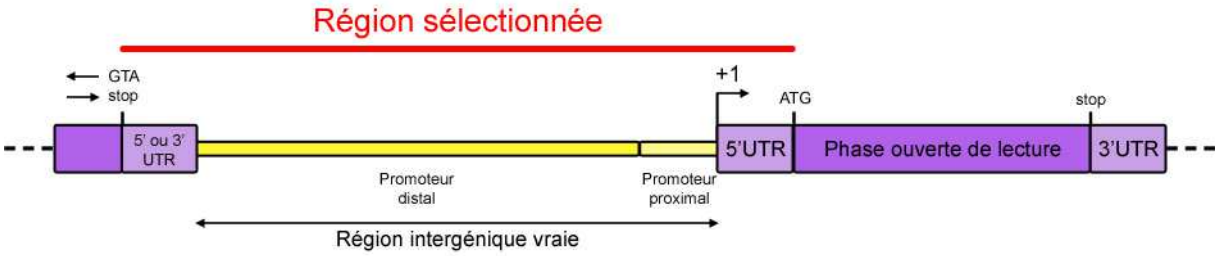


Figure 21. Schéma de la « région intergénique » sélectionnée pour la suite de l'étude.
« +1 » représente le site d'initiation de la transcription. « 5' et 3' UTR » représentent les régions qui seront transcrites mais non traduites. ATG et stop sont les codons respectivement d'initiation et de terminaison de la traduction d'un ARNm.

II.2 - Construction d'une bibliothèque de séquences promotrices

Notre étude de la transcription des gènes de *Plasmodium falciparum* se fait lors de la phase érythrocytaire parce que la culture du parasite dans les globules rouges est facile, mais aussi parce que c'est lors de la phase érythrocytaire qu'a lieu le passage de la forme asexuée du parasite (prolifération intense) à sa forme sexuée (gamétocytogénèse). En se basant sur ces observations, avant le séquençage complet du génome, les différents gènes étudiés ont été sélectionnés dans la littérature ; en effet, celle-ci rapportait depuis quelques années l'existence de transcrits préférentiellement exprimés lors du stade asexué ou sexué. Mais ceci ne signifiait pas que les ARNm étaient systématiquement traduits en protéines. Nous avions donc établi une liste de gènes connus et clonés, même si leur fonction précise n'était pas tout à fait établie alors [36].

Avec le séquençage complet du génome de *Plasmodium falciparum* et les avancées dans la technologie des puces à ADN pour étudier le transcriptome complet [39, 224] ou ciblé [320] du parasite lors de la phase érythrocytaire, nous avons élargi notre champ d'action aux promoteurs de tous les gènes du parasite impliqués dans nos recherche. Une bibliothèque de séquences a donc été mise au point. Elle regroupe les séquences des « régions intergéniques » du génome qui ont une longueur supérieure à 100 nucléotides et inférieure à 10 000 nucléotides [6] : elle est donc composée de 5 266 séquences.

II.3 - Recherche d'éléments de régulation

Pour l'identification des différents éléments de régulation, trois types de séquences ont été considérés : les séquences intergéniques de *P. falciparum* et les phases ouvertes de lecture qui leur sont associées, ainsi que des séquences aléatoires correspondant aux séquences intergéniques mélangées, qui ont donc gardé la même composition en bases et qui ont été obtenues avec le programme 'shuffleseq' du package EMBOSS [322]. Les différents motifs

[L] Les séquences d'une longueur supérieure à 10 000 nucléotides correspondent pour la plupart à des régions subtélomériques des chromosomes qui sont des séquences très biaisées : elles sont composées d'éléments répétés conservés d'un chromosome à l'autre.

ont été recherchés dans les séquences données ainsi que dans les séquences inverse-complémentaires.

Eléments de régulation connus chez les autres eucaryotes

Différents éléments de régulation connus chez les autres eucaryotes ont été recherchés, avec le programme PATTERNn [62]. Pour éviter toute redondance dans les résultats, les motifs recherchés correspondent à des consensus faits à partir d'éléments de régulation présents dans la partie MATRIX de la base de données TRANSFAC® (Figure 22) :

- pour la boîte TATA : TATAAA[at][agt],
- pour la boîte CCAAT : [agc][agt]CCAAT[cg][ag],
- pour la boîte GC : GGGGCGGGG[ct],
- pour le site de fixation des protéines Myb : [ct]AAC[acgt]G[act][act],
- pour le site de fixation des protéines C/EBP : T[gt][agt][cg]G[act]AA.
-

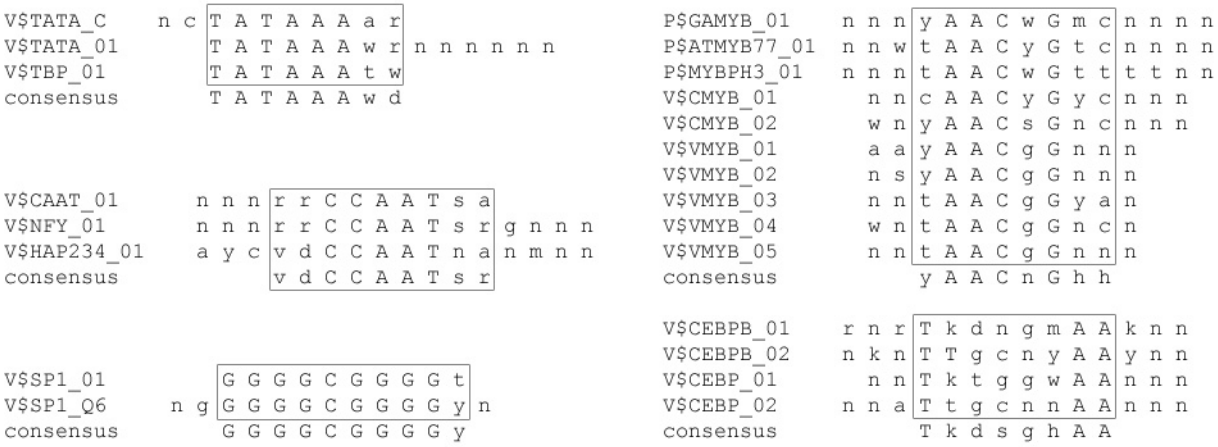


Figure 22. Eléments de régulation recherchés dans les séquences intergéniques de *P. falciparum*.
Pour chaque type d'éléments de régulation, un consensus a été fait à partir de la description de ces différents éléments dans la base de données TRANSFAC®. Le code de ces éléments comme par V ou par P ce qui indique la provenance des motifs : vertébrés ou plantes. Les nucléotides présents dans toutes les entrées de TRANSFAC® sont indiqués en majuscule. Les nucléotides sont codés de la manière suivante : M pour A ou C ; R pour A ou G ; W pour A ou T ; S pour C ou G ; Y pour C ou T ; K pour G ou T ; V pour A, C ou G ; H pour A, C ou T ; D pour A, G ou T ; N pour A, C, G ou T.

Ces éléments ont été recherchés soit indépendamment les uns des autres, soit en les combinant ensemble pour former ce que l'on appelle un module de régulation :

- CCAAT / TATA, la boîte TATA étant séparée de la boîte TATA de 5 à 100 nucléotides,

- *myb* / *c/ebp* ou *c/ebp* / *myb*, les motifs étant séparés les uns des autres par 100 nucléotides au maximum.

Eléments de régulation spécifiques de P. falciparum

Des éléments de régulation spécifiques de *P. falciparum* ont été recherchés dans différents groupes de promoteurs avec plusieurs programmes :

- dans le cas de AlignACE [177, 330], les motifs recherchés ont une taille comprise entre 6 et 8 nucléotides et sont attendus en 1 à 4 exemplaires sur chaque séquence ; le « contexte » est fixé à 13% de G+C.
- dans le cas de MEME (Multiple Em for Motif Elicitation) [16], les motifs recherchés ont une longueur comprise entre 6 et 8 nucléotides.
- dans le cas de GIBBS Motif Sampler [223], les motifs recherchés ont des tailles de 6 et 8 nucléotides.

Quand les motifs recherchés sont trop longs, tous les programmes identifient des motifs correspondant à des stretches de A, de T ou de A+T. C'est pourquoi la limite est à 8 nucléotides.

II.4 - Identification et annotation de facteurs de transcription

Une recherche par mots-clés et par homologie de séquences dans les bases de données SwissProt [34], PIR (Protein Information Resource) [409] & UniProt (Universal Protein Resource) [9, 17] *via* les outils SRS (Sequence Retrieval System) [108] et BLAST2 (Basic Local Alignment Search Tool) [7, 194, 195] ainsi qu'une recherche dans la base de données TRANSFAC® [255, 404] a permis de collecter bon nombre de séquences protéiques homologues au facteur de transcription recherché. Sur toutes ces séquences, le domaine caractérisant la famille de protéines, le plus souvent le domaine de liaison à l'ADN, a été identifié grâce à MotifScan [294]. Puis ces séquences ont été alignées avec ClustalW [64, 375] et MultAlin [64, 66], de façon à obtenir une séquence consensus du domaine caractéristique. Cette séquence consensus a alors été utilisée comme requête pour interroger, grâce aux outils TBLASTN et BLASTP [7, 194, 195], les bases de données de *Plasmodium falciparum*, présentes au début de mon travail sur le site d'Infobiogen sous forme de contigs et de shotguns. Une

fois le domaine caractéristique identifié dans le génome de *P. falciparum*, une recherche de phase ouverte de lecture a été effectuée manuellement.

Le domaine caractéristique de chaque séquence plasmodiale a ensuite été utilisé comme requête pour interroger de nouveau les bases de données de *Plasmodium*, au fur et à mesure du séquençage, de façon à identifier de nouveaux facteurs de la même famille.

C'est ainsi qu'ont pu être identifiés 4 facteurs structuraux de la famille HMG, 3 facteurs NF-Y se fixant à la boîte CCAAT, 3 facteurs dans la famille Myb et un facteur qui pourrait être un co-facteur de l'ARN polymérase III, le facteur TFIIIA.

Chacune des phases ouvertes de lecture identifiées a alors subi un examen de passage : tout d'abord, j'ai utilisé MotifScan [294] pour vérifier la présence du domaine caractéristique, mais aussi pour découvrir de nouveaux domaines, parmi lesquels des signaux de localisation nucléaire. En effet, comme les facteurs de transcription agissent dans le noyau de la cellule, ils devraient donc contenir dans leur séquence un signal qui indique à la cellule qu'après la traduction de l'ARNm en protéine, cette dernière devra être dirigée vers le noyau. Une deuxième vérification est effectuée grâce à PSORT [271] qui prédit la localisation de la protéine dans la cellule.

Une fois que le génome nucléaire du clone 3D7 a été entièrement séquencé et annoté en octobre 2002, j'ai pu vérifier la véracité des résultats et ainsi connaître le numéro d'accession dans PlasmoDB de chacun de ces facteurs et donc avoir accès à de nombreux renseignements répertoriés dans cette base de données.

II.5 - Phylogénie

Deux études phylogénétiques ont été menées lors de ce travail : l'une avec les facteurs HMGB et l'autre avec les facteurs NF-Y. La même méthode a été utilisée pour les deux études. Des séquences sont extraites des banques de données et leur motif caractéristique identifié par MotifScan [294].

Etude des facteurs HMGB

Dans l'étude faite par Stéphan Soullier et ses collaborateurs [353], une centaine de séquences HMG ont été utilisées auxquelles j'ai ajouté pour cette étude (Annexes I.1 & I.2) :

- les 4 séquences HMGB de *P. falciparum* (PlasmoDB : PFL0145c, MAL8P1.72, PFL0290w et MAL13P1.290)
- les séquences orthologues de PfHMGB1 et PfHMGB2 trouvées par homologie de séquences dans les génomes en cours de séquençage de *P. vivax*, de *P. yoelii* et *P. berghei*, deux parasites infectant les rongeurs ainsi que dans le génome de *P. knowlesi*, parasite infectant les singes,
- la séquence NHP1 de *Babesia bovis*, qui est un apicomplexe comme *Plasmodium*.

Cent cinquante-neuf domaines 'HMG-box' ont été identifiés par MotifScan (Pfam : PF00505) puis alignés avec ClustalW [375], en demandant le format PHYLIP comme format de sortie. Pour les protéines contenant plus d'un domaine 'HMG-box', chaque domaine a été traité séparément des autres. Les régions de l'alignement qui étaient plus ou moins bien alignées, comme les extrémités N- et C-terminales, ainsi que de longues insertions présentes dans seulement une ou deux séquences très proches, ont été exclues de l'étude. L'alignement initial des 159 domaines contenait 79 sites, tous variables. Après l'exclusion des insertions, l'alignement contenait 71 sites informatifs (Annexe I.3).

Etude des facteurs NF-Y

Aux séquences utilisées pour faire l'alignement présenté dans l'article de Christophe Romier et ses collaborateurs [326], j'ai ajouté les séquences identifiées comme étant des sous-unités NF-YB et NF-YC du facteur hétérotrimérique NF-Y (PlasmoDB : PF11_0477, PF13_0043 et PF14_0374).

Les dix-sept séquences utilisées pour l'analyse phylogénétique (Annexes I.1 & I.4) contenaient en plus du domaine 'CBFD_NFYB_HMF' identifié par MotifScan (Pfam : PF00808) des séquences conservées de part et d'autre du domaine caractéristique. Ces séquences ont été alignées avec ClustalW [375] ; l'alignement, édité sous le format PHYLIP, contenait 97 sites dont 80 sont informatifs (Annexe I.5)

Pour la partie phylogénie de ce travail, divers programmes appartenant à la version 3.63 du package PHYLIP (PHYLogeny Inference Package) ont été utilisés [113]. Ce package inclut des méthodes basées sur les matrices de distance, la parcimonie, la vraisemblance.

Dans les deux cas, pour construire les arbres phylogénétiques et tester la robustesse des branches, 1 000 alignements sont générés avec le programme SEQBOOT. La méthode utilisée ici pour le rééchantillonnage est le bootstrap : les nouveaux alignements créés ont la même longueur que l'alignement de départ et ont été générés par tirage aléatoire avec remise des sites de l'alignement de départ [112]. De cette façon, dans les alignements aléatoires créés, un même site peut apparaître plusieurs fois.

Le programme PROTDIST créé une matrice des distances pour chacun des alignements avec le modèle de Jones-Taylor-Thornton ou JTT [185]. Ensuite le programme NEIGHBOR construit un arbre non enraciné à partir de chaque matrice de distances soit avec la méthode de Neighbor-joining de Saitou & Nei [332], soit avec la méthode UPGMA (Unweighted Pairwise Group Method Average) [348] [7]. Les 1 000 arbres sont comparés et le programme CONSENSE créé un arbre consensus dans lequel les branches présentes dans moins de 600 arbres sont fusionnées. L'arbre consensus est visualisé grâce aux programmes DRAWGRAM ou DRAWTREE.

L'expérience a été répliquée quatre fois en utilisant chaque fois une graine différente pour la génération des nombres aléatoires.

[7] La méthode UPGMA est utilisée pour reconstruire des arbres phylogénétiques si les séquences ne sont pas trop divergentes. C'est une méthode hiérarchique ascendante ou méthode agrégative. UPGMA utilise un algorithme de regroupement séquentiel dans lequel les relations sont identifiées dans l'ordre de leur similitude et la reconstruction de l'arbre se fait pas à pas grâce à cet ordre. Il y a d'abord identification des deux séquences les plus proches et ce groupe est ensuite traité comme un tout, puis on recherche la séquence la plus proche et ainsi de suite jusqu'à ce qu'il n'y ait plus que deux groupes.
Quant à la méthode Neighbor-joining, elle tente de corriger la méthode UPGMA afin d'autoriser un taux de mutation différent sur les branches. Les données initiales permettent de construire une matrice qui donne un arbre en étoile. Cette matrice des distances est ensuite corrigée afin de prendre en compte la divergence moyenne de chacune des séquences avec les autres. Le programme recherche les deux plus proches voisins i et j qui, une fois réunis, minimiseront la longueur totale de l'arbre. La matrice des distances est alors recalculée en considérant le groupe (i, j) comme indissociable. Le processus est réitéré jusqu'à ce qu'il n'y ait plus que deux groupes.

II.6 - Modélisation par homologie

La modélisation par homologie est basée sur l'observation que des protéines de séquences suffisamment similaires ont généralement des repliements voisins. En effet, les protéines homologues ont pour caractéristique d'avoir des structures secondaires et tertiaires relativement bien conservées même si au niveau de la séquence en acides aminés, cette conservation est beaucoup moins importante. La méthode consiste donc à prédire la structure tridimensionnelle d'une séquence cible en utilisant une séquence « support » dont la structure tridimensionnelle est connue car elle a été obtenue par cristallographie ou spectroscopie RMN. Ici les séquences cibles sont les séquences des facteurs de transcription. Tout réside donc dans le choix de la séquence support car du taux d'identité entre la séquence cible et la séquence support va dépendre la qualité de la méthode, la limite inférieure se situant vers 25% d'identité.

Deux méthodes ont été utilisées pour modéliser les différents facteurs de transcription. La première méthode est une méthode « tout automatique ». Elle consiste à soumettre une séquence protéique sur un serveur, en l'occurrence le méta-serveur @TOME [73, 95, 178, 184, 200, 260] et à « attendre la réponse » : la séquence support considérée comme étant la meilleure est celle qui a le plus faible score TITO [217] parmi les scores négatifs. Une fois la séquence support choisie, il suffit de faire tourner le programme TITO pour avoir un modèle de la structure tridimensionnelle de la séquence cible.

La deuxième est une méthode un peu plus manuelle : tout d'abord il faut interroger la base de données PDB [28] grâce à Blast2P (sur le site de l'EBI), la séquence requête étant soit la séquence complète du facteur de transcription, soit la séquence de son domaine de liaison à l'ADN. Les principales séquences sélectionnées par le BlastP sont alors alignées grâce à DBClustal [376]. Ensuite les fichiers PDB qui ont été sélectionnés sont étudiés en détail : ces fichiers contiennent-ils une ou plusieurs protéines ? ces protéines sont-elles entières ou s'agit-il uniquement de certains domaines ? ces protéines présentent-elles des mutations ? de quel organisme ces protéines sont-elles issues ? les structures des protéines ont-elles été obtenues par cristallographie ou par spectroscopie RMN ? s'il s'agit de cristallographie, quelle est la résolution du cristal ? s'il s'agit de spectroscopie RMN, quelle est la meilleure structure parmi toutes celles obtenues ? Toutes ces questions permettent de choisir la

meilleure séquence support, c'est-à-dire la séquence qui présente un taux d'identité avec la séquence cible parmi les plus élevés, qui est la plus longue possible et qui ne présente pas de mutations.

Une fois cette séquence support choisie, il faut identifier tous les domaines ou résidus importants grâce aux bases de données de motifs comme PROSITE [179], Pfam [22], PDBSum [222] ou CATH [290, 291, 300] et à la bibliographie. Ensuite, il est nécessaire de faire une prédiction des structures secondaires de la séquence cible avec CONSENSUS [87] et de la comparer à la structure tridimensionnelle de la protéine support. S'il y a une incohérence entre les deux, il faut vérifier ce que donne la prédiction de structures secondaires de la protéine support.

Lorsque toutes ces informations sont réunies, les deux séquences doivent être alignées grâce à des outils d'alignement 2 à 2 (LFASTA & LALIGN) [57, 176], d'alignement multiple (DiAlign2, ClustalW, DBClustal & MultAlin) [66, 262, 375, 376] et doivent aussi être alignées en fonction de leurs structures secondaires (prédites ou réelles). Les 3 alignements sont alors comparés de façon à faire un alignement consensus optimal qui ne contiendra, si possible, pas de gap au niveau des structures secondaires prédites et sur lequel seront rajoutées les informations biologiques comme les résidus en contact avec l'ADN, les résidus impliqués dans un pont disulfure, etc. L'alignement élaboré servira au programme Modeller 7v7 (Figure 23) pour modéliser la séquence cible par homologie à la structure support.

Quand le nombre de modèles demandé a été généré (100 modèles, dans ce cas), les fonctions objectives associées à chaque modèle sont analysées : plus la fonction objective est faible, meilleur est le modèle. Les structures modèles retenues par leur fonction objective, ainsi que la structure support, sont alors analysées :

- un alignement structural est fait avec l'outil CE (Combinatorial Extension) [342] pour connaître le degré de similitude des structures,
- la qualité des structures est vérifiée avec les programmes Verify3D [37, 102, 243] et ProSa2003 (Protein Structure Analysis) [345], et SwissPDB Viewer [154] fournit un diagramme de Ramachandran ainsi qu'une liste des acides aminés trop proches les uns des autres,
- la conformation des chaînes latérales est vérifiée avec les outils SCit [136] et SCWRL 3.0 [52].

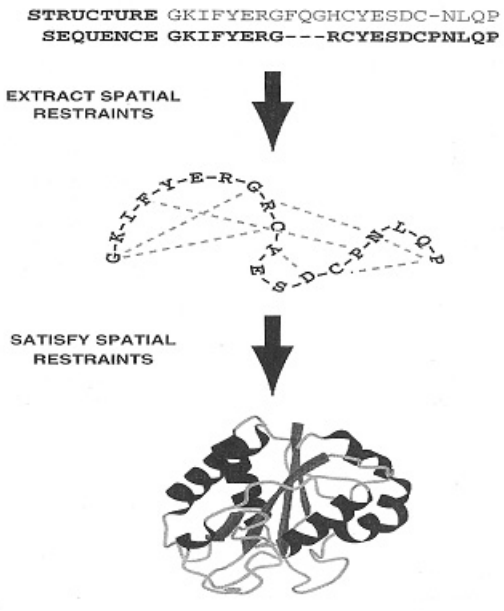


Figure 23. Construction d'une structure modèle par Modeller.
Tout d'abord, les contraintes spatiales, sous forme de distances atomiques et d'angles dièdres, sont extraites de la (ou des) structure(s) support(s). L'alignement est utilisé pour déterminer les résidus équivalents entre la cible et le support. Les contraintes servent à calculer une fonction objective. Finalement, la structure modèle de la séquence cible est optimisée jusqu'à ce qu'elle satisfasse au mieux les contraintes spatiales. Image tirée de Marti-Renom *et al.* [252]

Les différentes structures modèles sont alors alignées sur la structure support grâce à l'option 'Iterative Magic Fit' du logiciel SwissPDB Viewer (v3.7 SP5) ou grâce au logiciel ProFit [250], selon ce que l'on veut faire : en effet, avec SwissPDB Viewer, il est possible de superposer des structures ne comportant pas le même nombre d'acides aminés, ce qui est impossible avec ProFit. La qualité des modèles peut alors être aussi analysée visuellement avec le logiciel PyMol 0.97 [86].

RESULTATS

LES ELEMENTS DE REGULATION

Les éléments de régulation sont des séquences d'ADN d'une longueur souvent comprise entre quatre et dix nucléotides et sont répartis dans le promoteur sur un brin ou sur son complémentaire. Il existe deux types d'éléments de régulation (Figure 9) :

- des éléments de régulation dits « constitutifs » : on les retrouve, chez les eucaryotes, dans pratiquement tous les promoteurs proximaux,
- des éléments dits « spécifiques » : ils sont situés dans le promoteur proximal comme dans le promoteur distal et permettent la fixation de facteurs de transcription qui leur sont propres et qui agiront en synergie avec la machinerie basale de transcription pour moduler, positivement ou négativement, le niveau de transcription.

Ces différents éléments de régulation ont été recherchés soit dans tous les promoteurs de *Plasmodium*, soit dans des groupes de promoteurs formés parce que les gènes correspondants ont des points communs : soit ils ont le même profil d'expression, soit leur expression est modifiée lorsque l'on traite les cellules. Les résultats seront donc présentés dans l'ordre suivant : tout d'abord, ce qui concerne les éléments constitutifs et ensuite ce qui concerne les éléments spécifiques, qu'ils soient connus chez les autres eucaryotes ou bien spécifiques de *P. falciparum*.

I - Les éléments de régulation constitutifs

Les promoteurs de certains gènes de *P. falciparum* ont été moléculairement étudiés et leurs sites d'initiation de la transcription identifiés (Tableau 4) par diverses méthodes moléculaires : extension d'amorce, cartographie à la nucléase S1, protection contre l'action des ribonucléases, technique RACE (rapid amplification of cDNA ends). Ces promoteurs ont été utilisés pour tester différents outils disponibles sur la toile (NNPP [319], Promoter 2.0 [208], PromoterScan [308], TSSG, TSSW et TSSP [350]) qui permettent de prédire la position des sites d'initiation de la transcription dans le but de savoir s'ils étaient adaptés à un génome aussi particulier que celui du parasite.

Tableau 4. Promoteurs ayant servi de test pour les programmes de prédiction des sites d'initiation de la transcription.

Gène	N° accession PlasmoDB	RI	Position du SIT
pcna	PF13_0328	3457	-960 (*)
hsp86	PF07_0029	2286	-650
gbp130	PF10_0159	3141	-985
kahrp	PFB0100c	3767	-849
calmoduline	PF14_0323	3106	-62
pfs16	PFD0310w	3994	-175
pfs25	PF10_0303	774	-267 (*)

La colonne 'RI' indique la longueur de la région intergénique située en amont de chaque gène. La colonne 'Position du SIT' indique la position du site d'initiation de la transcription par rapport à l'ATG initiateur. L'astérisque (*) indique le site d'initiation de la transcription majeur quand plusieurs sites d'initiation de la transcription ont été identifiés. Les données sont issues des articles suivants : *pcna* [172], *hsp86* [360], *gbp130* [220], *kahrp* [219], *calmoduline* [324], *pfs16* et *pfs25* [84].

Dans les sept promoteurs choisis pour le test, aucun des sites d'initiation de la transcription identifiés par approche moléculaire n'a été prédit correctement et ce, par aucun des programmes utilisés (données non montrées). Ce résultat n'est pas surprenant outre mesure. En effet, ces programmes de prédiction ont été mis au point pour l'étude du génome humain (et par extension des génomes de primates) ou pour les génomes de ce que l'on appelle les organismes modèles, à savoir la souris, la drosophile ou encore la plante *Arabidopsis thaliana*. Ces génomes sont très éloignés de celui de *P. falciparum* de par la taille, la densité de gènes et surtout la composition en bases (Figure 4). De plus, ces programmes ne sont pas très fiables : J. W. Fickett & A. G. Hartzigeorgiou les ont testés avec un groupe de promoteurs de mammifères dont le site d'initiation de la transcription avait été cartographié expérimentalement et selon le programme utilisé, entre 13 et 54% des sites d'initiation de la transcription seulement ont été correctement prédits [115]. Pour la suite de l'étude des promoteurs de *P. falciparum*, nous avons donc utilisé les régions intergéniques complètes sans nous préoccuper de la position du site d'initiation de la transcription (Figure 21) mais en gardant évidemment à l'esprit que les régions transcrites mais non traduites (UTR) inclues dans les séquences intergéniques peuvent gêner l'interprétation des résultats.

Le promoteur proximal est situé juste en amont de l'unité de transcription et comporte, dans la plupart des gènes, des motifs dits « constitutifs », parmi lesquels la boîte TATA où

vient se fixer la TBP, la boîte GC qui est reconnue par le facteur Sp1 et la boîte CCAAT sur laquelle se fixe, entre autres, le facteur NF-Y (Figure 9). La présence et la fréquence de ces boîtes ont été étudiées dans les promoteurs de tous les gènes de classe II de *P. falciparum* et comparées aux résultats obtenus avec les séquences codantes de *P. falciparum* et des séquences aléatoires ayant la même composition en bases que les régions intergéniques (Tableau 5).

A cause de l'extrême richesse en A+T du génome de *Plasmodium falciparum*, qui est plus particulièrement prononcée dans les régions intergéniques (Tableau 3), on trouve, dans les régions se situant en amont des gènes, un très grand nombre de séquences ressemblant à la classique boîte TATA. En effet, 115 159 motifs correspondant au consensus TATAAA[at][agt] de la boîte TATA sont répartis dans 5 225 régions intergéniques (soit 99% des séquences), ce qui fait une moyenne d'environ 10 boîtes TATA pour 1 000 pb.

Tableau 5. Boîtes TATA et CCAAT identifiées dans tous les promoteurs de *P. falciparum*.

		Nb séquences	Nb motifs	Moyenne	Médiane
boîte TATA	(RI)	5225	115159	10,58	10,00
boîte TATA	(SC)	4789	49662	4,12	3,79
boîte TATA	(SA)	5226	115052	10,78	10,62
boîte CCAAT	(RI)	266	275	0,47	0,41
boîte CCAAT	(SC)	1068	1411	0,64	0,42
boîte CCAAT	(SA)	264	280	0,55	0,47
module CCAAT-TATA	(RI)	121	192	0,75	0,54
module CCAAT-TATA	(SC)	276	343	0,46	0,29
module CCAAT-TATA	(SA)	131	194	0,73	0,50

Les boîtes TATA et CCAAT ainsi que leur association en module de moins de 100 pb ont été recherchées dans les régions intergéniques (RI), dans les séquences codantes associées (SC) et dans des séquences aléatoires de même composition que les régions intergéniques (SA). Le nombre de séquences présentant ces boîtes ainsi que le nombre total de motifs mis en évidence sont indiqués. La moyenne et la médiane correspondent à la moyenne et la médiane du nombre d'éléments ou de modules de régulation pour une séquence de 1 000 pb.

Le biais de composition est responsable de ce nombre élevé de boîtes TATA mais il est aussi responsable du fait que les régions intergéniques ne contiennent aucun motif correspondant à la boîte GC. On pouvait s'attendre à ce résultat car le consensus de la boîte GC est composé de 8 guanines séparées en deux groupes par une cytosine. Le parasite pourrait donc s'être adapté et avoir remplacé le couple « boîte GC- facteur Sp1 » que l'on rencontre chez beaucoup d'eucaryotes par un autre couple avec un élément de régulation plus en accord avec le biais de composition des régions intergéniques et un facteur de

transcription qui lui correspond. En effet, non seulement il n'existe pas de boîte GC canonique dans le génome de *Plasmodium falciparum* mais nous n'avons pas non plus été en mesure d'annoter un facteur Sp1 correspondant lors de nos recherches (voir p. 152).

Entre ces deux extrêmes se situe la boîte CCAAT. Il existe 275 boîtes CCAAT réparties dans 266 promoteurs. Même si le consensus [agc][agt]CCAAT[cg][ag] qui a été utilisé pour rechercher les boîtes CCAAT dans les promoteurs est riche en A+T, seuls 5% des séquences étudiées pourraient contenir une boîte CCAAT putative.

Quand le module composé d'une boîte CCAAT suivie par une boîte TATA dans les 100 nucléotides situés en aval est recherché dans les séquences intergéniques, les résultats sont un peu étonnants : en effet, comme les séquences intergéniques renferment un nombre très élevé de boîtes TATA (en moyenne 10 boîtes pour 1 000 pb), on aurait pu penser que toutes les boîtes CCAAT étaient suivies de près par une boîte TATA. Or sur les 266 séquences promotrices possédant au moins une boîte CCAAT, seules 121 possèdent au moins un module de régulation CCAAT-TATA. De plus, quand une séquence possède plusieurs modules, il s'agit très souvent d'une même boîte CCAAT associée à plusieurs boîtes TATA situées à moins de 100 nucléotides en aval.

Que l'on considère la boîte TATA, la boîte CCAAT ou l'association en module de régulation de ces deux boîtes, les résultats obtenus avec les séquences intergéniques vraies et les séquences aléatoires ne sont pas très différents (Tableau 5). En revanche, les résultats obtenus avec les séquences codantes, donc les séquences où les boîtes TATA et CCAAT putatives sont censées être inactives, les résultats sont très différents : les boîtes TATA sont très nettement sous-représentées par rapport aux deux autres types de séquences tandis que les boîtes CCAAT et l'association des boîtes CCAAT et TATA sont surreprésentées. Ces résultats peuvent s'expliquer par la différence de composition des régions codantes et des régions non codantes : en effet, le pourcentage en A+T est moins prononcé dans les séquences codantes, d'où une diminution du nombre des boîtes TATA et une augmentation du nombre des boîtes CCAAT. L'augmentation du nombre de modules de régulation CCAAT-TATA dans les séquences codantes est en rapport avec le nombre supérieur de boîtes CCAAT, car même si les séquences codantes ont une composition en A+T inférieure à celle des séquences non codantes, le nombre de boîtes TATA reste tout de même important.

Même si la boîte CCAAT est très nettement moins représentée que la boîte TATA dans les séquences intergéniques vraies et l'association de ces deux boîtes retrouvées dans uniquement 5% des promoteurs, il se pourrait donc que leur apparition dans les promoteurs soit le fait du hasard.

On peut alors se demander comment la machinerie basale de transcription du parasite peut faire la différence entre des séquences qui représentent de vrais signaux indiquant où doit notamment se fixer la TBP et des séquences qui ne sont que des conséquences du biais de composition. Au jour d'aujourd'hui, il existe très peu de gènes de *Plasmodium falciparum* dont les promoteurs ont été étudiés en détail et validés moléculairement ; de plus, quand les boîtes TATA et CCAAT sont mentionnées, c'est uniquement parce qu'elles ont été annotées par homologie aux consensus connus et non parce qu'elles ont été caractérisées moléculairement. Les caractéristiques des vrais sites de fixation de la TBP ou du facteur NF-Y ne sont pas connues chez *Plasmodium*. Ces sites de fixation peuvent légèrement différer de ce que l'on connaît chez les autres eucaryotes à cause du biais de composition. Sans oublier que les séquences flanquantes de ces éléments de régulation apportent une information supplémentaire et sont donc importantes pour la reconnaissance du motif et la fixation des protéines de régulation.

II - Les éléments de régulation spécifiques

Au cours de ce travail, nous avons essayé d'annoter, chez *Plasmodium,* des facteurs de transcription par homologie avec des facteurs connus chez d'autres eucaryotes et appartenant aux différentes familles définies par TRANSFAC®. On peut donc penser que les facteurs plasmodiaux identifiés se fixent sur des séquences similaires à celles que l'on rencontre chez les autres eucaryotes. C'est pourquoi nous avons recherché dans les promoteurs de tous les gènes de *P. falciparum* les sites de fixation de la protéine PfMyb1 que nous avons annotée et étudiée moléculairement.

En même temps, il semblerait logique de penser que la richesse du génome en A+T a obligé le parasite à s'adapter et donc à développer de nouveaux éléments de régulation qui seront les sites de fixation de facteurs de transcription spécifiques du parasite. Des groupes

de promoteurs définis à partir de résultats d'expériences biologiques ont donc été étudiés pour mettre en évidence des éléments de régulation propres à *P. falciparum*.

II.1 - Le site de fixation de la protéine Myb

Nous avons identifié chez *P. falciparum* un facteur de transcription de la famille Myb : PfMyb1 (voir p. 157). Ce facteur se fixe sur des séquences issues de promoteurs parasitaires et répondant au consensus [ct]AAC[acgt]G[act][act] . Nous avons voulu analyser la présence de ce motif dans les régions intergéniques de *P. falciparum*.

Des éléments de régulation de type *myb* ont été identifiés dans 50% des séquences intergéniques avec en moyenne moins d'un élément de régulation pour 1 000 pb (Tableau 6). Cet élément de régulation semble être sous-représenté dans les régions intergéniques par comparaison à ce que l'on rencontre dans des séquences aléatoires de même composition que les régions intergéniques. En effet, 66% des ces séquences possèdent en moyenne un peu plus d'un élément de régulation de type *myb* pour 1 000 pb.

Tableau 6. Eléments de régulation de type *myb* et *c/ebp* identifiés.

		Nb séquences	Nb motifs	Moyenne	Médiane
élément de type *myb*	(RI)	2687	4799	0,86	0,70
élément de type *myb*	(SC)	3942	12885	1,70	1,27
élément de type *myb*	(SA)	3473	7675	1,10	0,90
élément de type *c/ebp*	(RI)	3290	7181	1,09	0,90
élément de type *c/ebp*	(SC)	3975	13798	1,72	1,36
élément de type *c/ebp*	(SA)	2920	6523	1,10	0,88
module *myb -c/ebp*	(RI)	264	338	0,64	0,48
module *myb -c/ebp*	(SC)	444	522	0,69	0,42
module *myb -c/ebp*	(SA)	331	485	0,74	0,55
module *c/ebp -myb*	(RI)	250	312	0,35	0,47
module *c/ebp -myb*	(SC)	253	299	0,71	0,44
module *c/ebp -myb*	(SA)	337	483	0,73	0,53

Les éléments de régulation de type *myb* et de type *c/ebp* ainsi que leur association en module de régulation de moins de 100 nucléotides ont été recherchés soit dans les régions intergéniques (RI) soit dans ces mêmes régions mais mélangées (RM). Le nombre de séquences présentant ces éléments ainsi que le nombre total de motifs mis en évidence sont indiqués. La moyenne et la médiane correspondent à la moyenne et la médiane du nombre d'éléments ou de modules de régulation pour une séquence de 1 000 pb.

Dans la littérature se trouvent des informations sur des gènes possédant dans leur promoteur des éléments de régulation de type *myb*. Il est connu que la protéine c-Myb peut

jouer un rôle direct dans le contrôle de gènes impliqués dans le cycle cellulaire. Elle fonctionne alors comme un facteur qui permet la progression G1→S [143] en influant sur les niveaux d'ARN de divers facteurs impliqués dans la synthèse d'ADN : la kinase $p34^{cdc2}$ [128], le facteur PCNA et l'ADN polymérase α [386]. Chez *P. falciparum*, les promoteurs de ces gènes (MAL13P1.279, PF13_0328 et PFD0590c) contiennent eux aussi des éléments de régulation de type *myb* putatifs.

Chez les eucaryotes, les promoteurs des gènes ciblés par les protéines c-Myb contiennent souvent des sites de liaison pour les membres du sous-groupe C/EBP qui appartiennent à la famille des régulateurs transcriptionnels à agrafe à leucines (Figure 15a). En effet des gènes spécifiques des monocytes comme *mim-1* [277], *tom-1A* [47] et le gène du lysozyme [276] ou encore les gènes de recombinaison *rag1* et *rag2* [121] contiennent des sites de liaison pour les deux facteurs c-Myb et C/EBP et leur expression est activée de manière coopérative par ces deux facteurs. C'est pourquoi des modules de régulation composés d'un élément de régulation de type *myb* et d'un élément de régulation de type *c/ebp* ont été recherchés. L'orientation des éléments sur l'un ou l'autre des brins d'ADN a été prise en compte ainsi que l'ordre d'apparition des deux éléments (Figure 24).

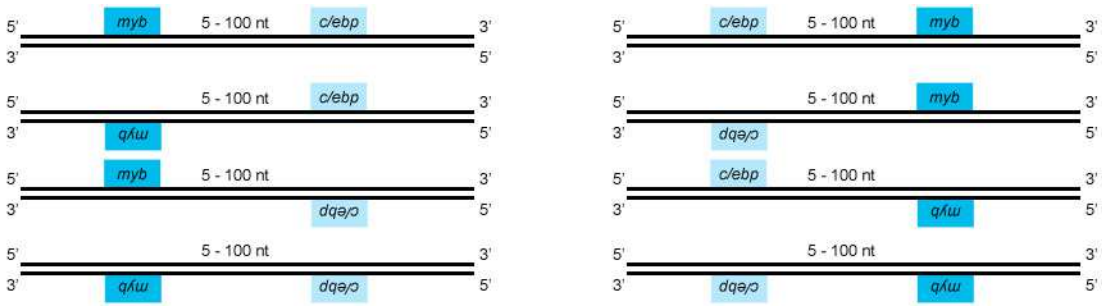


Figure 24. Modules de régulation composés d'éléments de type *myb* et *c/ebp*.
Les éléments de régulation ont été recherchés sur le brin codant comme sur son complémentaire, l'élément de régulation *myb* situé en amont de l'élément de régulation *c/ebp* et inversement. Les deux éléments de régulation sont séparés par 100 nucléotides au maximum.

L'élément de régulation de type *c/ebp* est retrouvé plus fréquemment dans les régions intergéniques que dans les séquences aléatoires mais avec la même moyenne. Il apparaît aussi plus souvent que l'élément de régulation de type *myb*, dans un plus grand nombre de séquences et à une moyenne plus élevée.

L'association des deux éléments de régulation à moins de 100 nucléotides de distance ne semble pas être due au hasard. En effet, qu'il s'agisse du module *myb-c/ebp* ou du module *c/ebp-myb*, il y a moins d'occurrences dans les séquences intergéniques que dans les séquences aléatoires. En ce qui concerne plus particulièrement le module *c/ebp-myb*, il y a, en moyenne, deux fois moins de modules dans une séquence intergénique que dans une séquence aléatoire.

Enfin, que l'on considère l'élément de régulation de type *myb*, l'élément de régulation de type *c/ebp* ou l'association des deux, il existe une forte différence entre les séquences intergéniques et donc non codantes et les séquences codantes (Tableau 6). En effet, les deux éléments pris indépendamment l'un de l'autre sont trouvés en moyenne à 1,7 motif pour 1 000 pb dans 75% des séquences codantes. Quant à l'association des deux éléments :

- dans le cas du module *myb-c/ebp*, elle est représentée dans un plus grand nombre de séquences (522 modules répartis dans 444 séquences codantes contre 338 modules dans 264 séquences intergéniques) ;
- dans le cas du module *c/ebp-myb*, elle apparaît en moyenne deux fois plus souvent dans les séquences codantes (0,71 module pour 1 000 pb contre 0,35 dans les régions intergéniques).

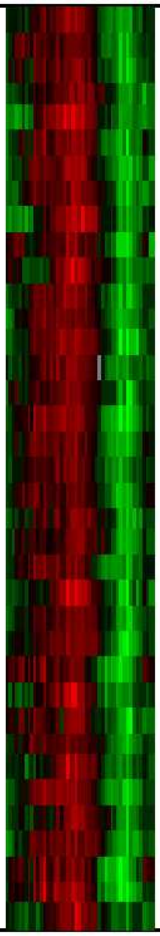
II.2 - Les promoteurs des gènes ayant le même profil d'expression que *pfmyb1* ont des motifs communs

Avec le séquençage complet du génome de *Plasmodium falciparum* et les avancées dans la technologie des puces à ADN, deux études globales d'expression des gènes ont été publiées en 2003 [39, 224]. L'une d'entre elles a été réalisée par l'équipe de Joseph L. DeRisi (UCSF, Californie) et a consisté à suivre l'expression des gènes heure par heure tout au long du cycle érythrocytaire (48h) [39]. Cette étude a montré qu'environ 60% des gènes sont exprimés tout au long du cycle avec un pic maximal et un pic minimal d'expression. Sur le site Internet du laboratoire de Joseph L. DeRisi (http://malaria.ucsf.edu/), toutes les données de leurs puces à ADN sont accessibles.

Le gène *pfmyb1* a un pic d'expression maximal à 23h post-infection (stade trophozoïte) et un pic minimal à 40h post-infection (stade schizonte). L'amplitude entre ces deux pics est de

2,7. Les promoteurs des gènes ayant le même profil d'expression ont été sélectionnés selon les critères suivants : (i) pic maximal : 23h ± 1h et pic minimal : 40h ± 1h, (ii) amplitude entre les deux pics supérieure à 2 et inférieure à 3 et (iii) longueur du promoteur supérieure à 100 nucléotides. Ces critères ont permis de sélectionner 36 gènes ayant le même profil d'expression que *pfmyb1* (Tableau 7).

Tableau 7. Liste des gènes ayant le même profil d'expression que *pfmyb1*.

N° accession PlasmoDB	Longueur du promoteur	Profil d'expression	Heure maximale	Heure minimale	Amplitude (log2)	Annotation putative Fonction putative
PFB0545c	1797		22	40	2,3	ribosomal protein L7/L12, putative
PFC0850c	964		23	40	2,1	hypothetical protein
PFD0835c	825		24	40	2,4	hypothetical protein
PFE0290c	707		23	40	2,1	hypothetical protein
PFE0980c	3520		23	40	2,5	hypothetical protein
PFE1055c	866		24	40	2,4	hypothetical protein
MAL6P1.141	1542		23	41	2,6	hypothetical protein
PF07_0102	840		24	40	2,5	hypothetical protein
MAL8P1.157	2741		24	41	2,5	hypothetical protein
PF08_0084	1760		22	40	2,6	hypothetical protein
PFI0095c	2534		23	40	2,5	hypothetical protein
PFI0330c	473		22	41	2,5	hypothetical protein
PFI1385c	2207		23	40	2,1	hypothetical protein
PFI1650w	1733		23	41	2,8	DNA excision-repair helicase, putative
PFI1795c	2433		22	39	2,1	hypothetical protein
PF10_0173	1219		22	41	2,4	hypothetical protein
PF10_0214	4365		23	40	2,8	hypothetical protein
PF10_0366	2917		23	40	2,4	ADP/ATP transporter on adenylate translocase
PF11_0118	1278		22	39	2,1	hypothetical protein
PF11_0123	807		24	39	2,2	hypothetical protein
PF11_0292	1185		24	41	2,4	hypothetical protein
PFL0225c	855		24	40	2,4	hypothetical protein
PFL0380c	565		23	40	2,1	tRNA delta(2)-isopentenylpyrophosphate transferase
PFL0690c	535		23	40	2,8	hypothetical protein
PFL2030w	509		23	41	2,3	queuine tRNA ribosyltransferase, putative
PFL2395c	907		22	41	2,6	dimethyladenosine transferase, putative
PFL2480w	1019		23	39	2,4	hypothetical protein
PF13_0071	2743		23	40	2,5	hypothetical protein
PF13_0088	**2479**		**23**	**40**	**2,7**	**PfMyb1**
PF13_0142	1463		22	41	2,1	u6 snRNA-associated sm-like protein, putative
PF13_0183	871		22	40	2,2	hypothetical protein
PF13_0240	1507		22	41	2,5	aspartate carbamoyltransferase
PF13_0348	2507		24	41	2,4	PfRhop148, Rhoptry protein
PF14_0378	2781		22	41	2,2	triose-phosphate isomerase
PF14_0434	2697		24	39	2,6	hypothetical protein
PF14_0510	1800		23	40	2,9	hypothetical protein
PF14_0582	858		24	40	2,1	hypothetical protein

Dans la troisième colonne sont représentés les profils d'expression selon l'étude de l'équipe de Joseph L. DeRisi. Chaque ligne représente un gène et chaque colonne une heure du cycle érythrocytaire. Les couleurs rouge et vert représentent, respectivement, une sur- et une sous-expression par rapport à un niveau basal, la couleur noir un niveau équivalent au niveau basal. L'intensité est proportionnelle au niveau d'expression détecté.

Les résultats obtenus pour ces 37 régions intergéniques ont été comparés aux résultats obtenus pour 37 séquences aléatoires de même composition en bases que les régions intergéniques ainsi qu'aux résultats obtenus pour 37 régions intergéniques, prises au hasard

dans le génome de *P. falciparum* et qui gouvernent des gènes n'ayant pas du tout le même profil d'expression (Figure 25).

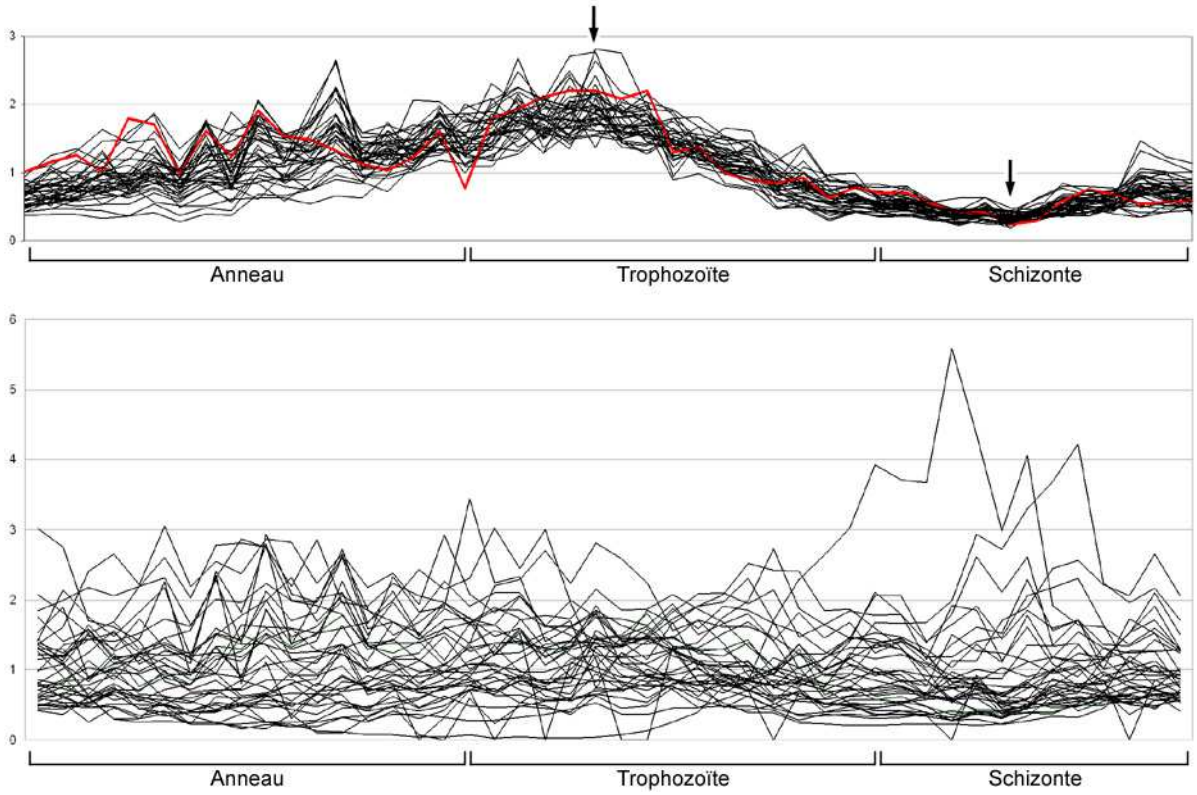


Figure 25. Profils d'expression.
(Haut) Profils d'expression lors des 48 heures du cycle érythrocytaire des 36 gènes (en noir) ayant le même profil d'expression que *pfmyb1* (en rouge). **(Bas)** Profils d'expression lors des 48 heures du cycle érythrocytaire des 37 gènes gouvernés par les régions intergéniques choisies au hasard dans le génome de *P. falciparum*.

Les programmes AlignACE et MEME ont mis en évidence deux types de motifs grâce aux promoteurs de ces 37 gènes (Figure 26a), le premier motif ayant aussi été prédit par le programme GIBBS Motif Sampler. Ces motifs n'apparaissent pas dans tous les promoteurs (Figure 26b), mais il semblerait qu'ils n'apparaissent pas au hasard. En effet, dans les séquences aléatoires, ces deux motifs apparaissent uniquement dans deux séquences.

Le motif 1 est représenté de la même manière dans les promoteurs des gènes ayant le même profil d'expression que pfmyb1 que dans les promoteurs pris au hasard (1 motif pour 1 000 pb dans une vingtaine de séquences). Il est aussi présent dans les séquences codantes mais moins souvent (0,75 motif pour 1 000 pb).

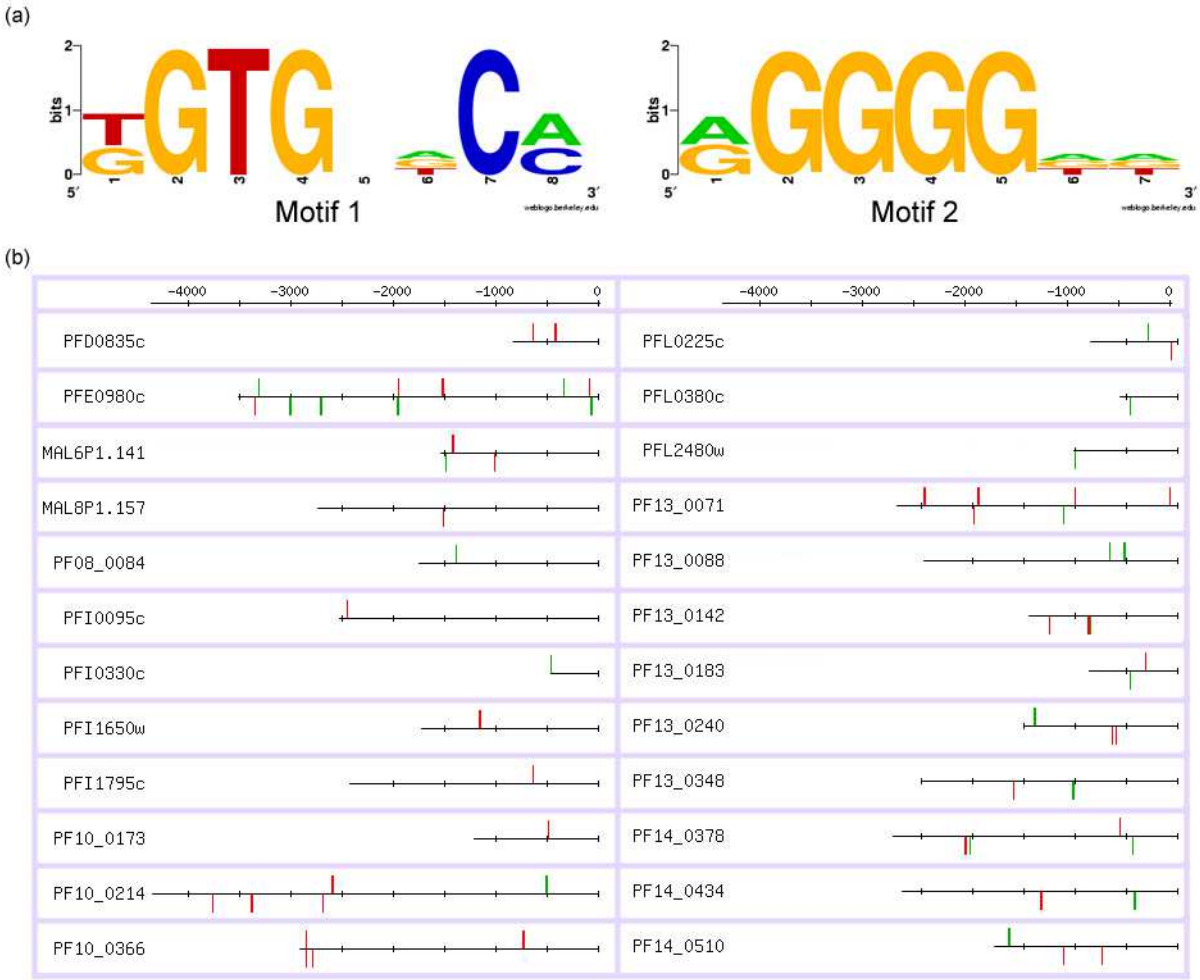


Figure 26. Deux motifs sont présents certains promoteurs gouvernant les gènes ayant le même profil d'expression que *pfmyb1*.
(a) Consensus des deux motifs identifiés par les programmes AlignACE, MEME et GIBBS Motif Sampler. A chaque position se trouvent les nucléotides possibles, une position vide indique que les quatre nucléotides sont acceptés. **(b)** Répartition des deux motifs dans certains promoteurs gouvernant les gènes ayant le même profil d'expression que *pfmyb1*. Les traits situés au dessus indiquent les motifs situés sur le brin codant et les traits en dessous, les motifs sur le brin complémentaire. En rouge, le motif 1 et en vert, le motif 2.

Le motif 2 est présent dans les promoteurs d'intérêt, dans les promoteurs pris au hasard ainsi que dans les séquences codantes mais à chaque fois avec des moyenne différente : 1 motif pour 1 000 pb dans les promoteurs contre 0,5 motif et 1,5 motif pour 1 000 pb pour les promoteurs choisis au hasard et les séquences codantes.

Quand un mésappariement est autorisé dans la recherche des motifs, le nombre d'occurrences augmente de façon considérable dans tous les promoteurs même si le promoteur du gène PFI1385c ne présente toujours aucun des deux motifs. Le motif 1 qui n'apparaissait que dans deux des séquences aléatoires est alors présent dans une trentaine de

séquences. En revanche, le motif 2, du fait de sa grande richesse en G, reste spécifique des séquences intergéniques vraies même avec un mésappariement.

II.3 - Les gènes dont l'expression est modifiée par un niveau diminué de *pfmyb1* ont des motifs partagés

Une culture de parasites a été traitée avec un double brin d'ARN de façon à inhiber la protéine PfMyb1 dans les parasites. Grâce à la puce à ADN du laboratoire ciblée sur le cycle cellulaire, la transduction du signal et la régulation de la transcription, les conséquences de cette inhibition sur l'expression de certains gènes ont pu être étudiées (Article 4). Il en ressort que huit gènes sont exprimés différemment (Tableau 8) : sept sont sous-exprimés (MAL6P1.248, MAL6P1.249, PF14_0224, PFI1105w, PFL1285c, PFL1885c, PFL2345c) et un seul est surexprimé (MAL13P1.279) (Figure 5, p. 35 de l'Article 4).

Tableau 8. Liste des gènes exprimés différemment lorsque la culture de parasites est traitée par un ARN double brin *pfmyb1*.

N° accession PlasmoDB	RI		Fonction
MAL6P1.248	5581	-	histone H3
MAL6P1.249	5581	-	histone H2A
PF14_0224	1046	-	PP1-like protein serine/threonine phosphatase
PFI1105w	1168	-	phosphoglycerate kinase
PFL1285c	2421	-	proliferating cell nuclear antigen, putative
PFL1885c	4035	-	calcium/calmodulin-dependent protein kinase 2, putative
PFL2345c	1144	-	TATA-binding protein homolog
MAL13P1.279	1061	+	cell division control protein 2 homolog

La colonne 'RI indique la longueur de la région intergénique propre à chaque gène. Les signes '-' et '+' indique si les gènes ont été sous-exprimés ou surexprimés suite au traitement de la culture.

Comme les gènes ont vu leur profil d'expression changer à cause de l'inhibition du facteur PfMyb1, les éléments de régulation de type *myb* ont été recherchés dans les promoteurs de ces gènes. Ces éléments de régulation ont été identifiés dans sept des huit promoteurs : seul le promoteur du gène codant la phosphoglycérate kinase (PFI1105w) ne possède pas de site de fixation canonique pour la protéine PfMyb1.

En plus des sites de fixation de la protéine Myb, des motifs partagés par ces huit régions intergéniques ont été recherchés. Quatre motifs ont été mis en évidence sur les promoteurs de ces gènes, par les différents programmes utilisés (Figure 27). Tous ces motifs ont en commun d'avoir principalement des guanines à des positions non ambiguës.

Les motifs 1 et 2 sont retrouvés dans les promoteurs de tous les gènes impliqués dans l'étude. Le motif 1 est représenté de manière assez homogène dans tous les promoteurs avec en moyenne 3,95 motifs pour 1 000 pb. Le motif 2, quant à lui, est représenté différemment selon les séquences : environ 3,6 fois pour 1 000 pb dans les promoteurs PFL1285c, PFI1105w, PFL2345c et PF14_0224 et MAL13P1.279) et environ 5,6 fois pour 1 000 pb dans les promoteurs MAL6P1.248, MAL6P1.249 et PFL1885c. Ces deux motifs sont surreprésentés par rapport à ce qui a été identifié dans des séquences aléatoires (respectivement 0,5 motif pour 1 000 pb et 2,15 motifs pour 1 000 pb). Mais ils sont sous-représentés par rapport aux résultats obtenus avec le groupe de promoteurs pris au hasard (respectivement 4,10 motifs pour 1 000 pb et 4,88 motifs pour 1 000 pb) et encore plus par rapport à ceux obtenus avec les séquences codantes (respectivement 7,18 motifs pour 1 000 pb et 6,69 motifs pour 1 000 pb).

Les deux autres motifs sont plus rares et ne sont pas présents dans toutes les séquences. Mais il semblerait qu'ils n'apparaissent pas au hasard. En effet, dans les séquences aléatoires, ces deux motifs n'apparaissent jamais.

Le motif 3 est identifié dans la moitié des promoteurs : dans les promoteurs PFL1885c et PFL1285c, il apparaît moins d'une fois pour 1 000 pb alors que dans les promoteurs PFI1105w et MAL13P1.279, il apparaît respectivement 2,6 et 2,8 fois pour 1 000 pb. Ce motif est aussi retrouvé dans la moitié des promoteurs choisis au hasard et la moitié des séquences codantes avec des moyennes équivalentes dans les trois groupes de séquences (entre 0,7 et 0,8).

Le motif 4 est absent des promoteurs PFL2345c et PF14_0224 et présent moins d'une fois pour 1 000 pb dans les autres séquences. Ce motif est quasiment absent des promoteurs pris au hasard et des séquences codantes car il est identifié respectivement dans deux et trois séquences uniquement.

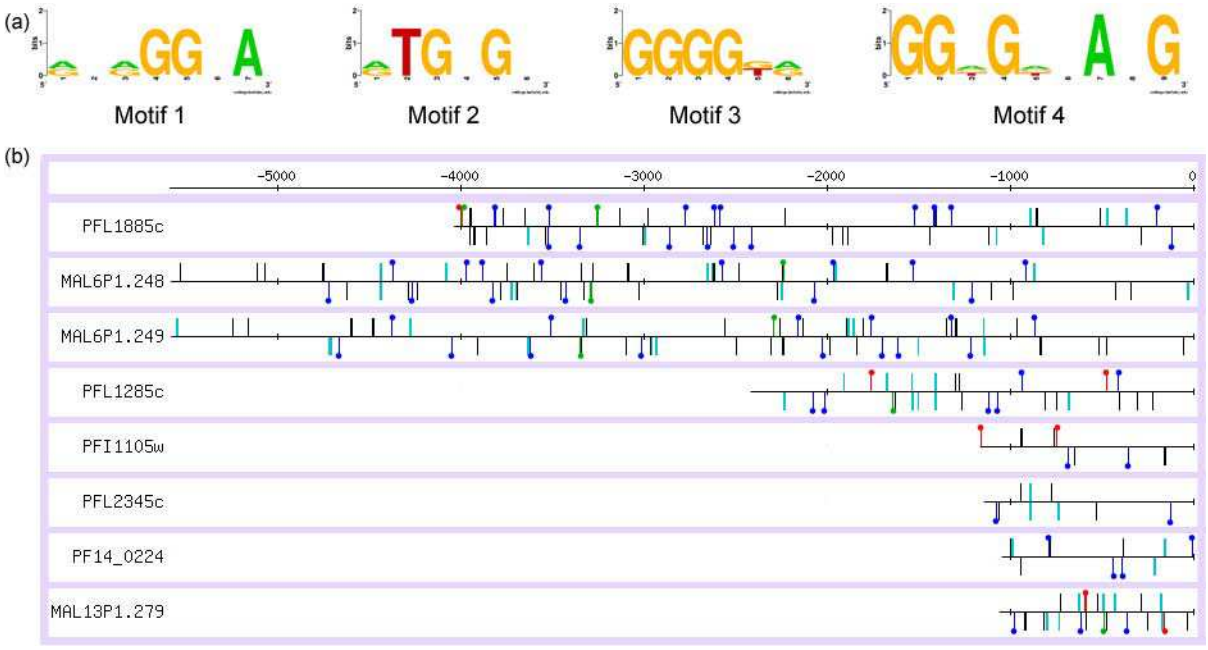


Figure 27. Quatre motifs sont partagés par les promoteurs des gènes dont l'expression est altérée quand le facteur de transcription PfMyb1 est inhibé.
(a) Consensus des quatre motifs identifiés par les programmes AlignACE et GIBBS Motif Sampler. A chaque position se trouvent les nucléotides possibles, une position vide indique que les quatre nucléotides sont acceptés. **(b)** Répartition des quatre motifs dans les promoteurs gouvernant les gènes dont l'expression a été modifiée à cause de l'inhibition du facteur PfMyb1. Les traits situés au dessus indiquent les motifs situés sur le brin codant et les traits en dessous, les motifs sur le brin complémentaire. En bleu ciel, l'élément de régulation de type *myb* ; en bleu, le motif 1 [ag][acgt][ag]GG[acgt]A ; en noir, le motif 2 [ag]TG[acgt]G[acgt] ; en rouge, le motif 3 GGGG[gt][ag] et en vert, le motif 4 GG[agt]G[agt][acgt]A[acgt]G. Les traits représentant les motifs 1, 3 et 4 ont été agrémentés d'un point de couleur pour faciliter la lecture de la figure.

Les gènes cibles de PfMyb1 sont exprimés différemment lorsque le facteur de transcription est inhibé : le gène MAL13P1.279 est surexprimé alors que les sept autres gènes sont sous-exprimés. Dans ces résultats, rien ne peut expliquer cette différence d'expression. Il n'existe aucun motif mis en évidence ici qui serait présent dans les promoteurs des sept gènes sous-exprimés et absent dans le promoteur du gène MAL13P1.279 ou présent dans les huit promoteurs mais dans des proportions différentes.

III - Discussion et perspectives

Plasmodium est un organisme tout à fait particulier car c'est l'organisme qui possède le génome le plus riche en A+T des organismes connus aujourd'hui (80,6% en moyenne). Cette richesse en A+T n'est pas répartie aléatoirement dans le génome. En effet, les régions non codantes, c'est-à-dire les introns et les régions intergéniques, peuvent posséder jusqu'à 90% de A+T. L'étude des promoteurs est donc rendue difficile par ce biais de composition.

Comme chez les autres eucaryotes, les promoteurs de *P. falciparum* possèdent une structure bipartite, à savoir un promoteur proximal qui permet la fixation de la machinerie basale de transcription et qui est donc responsable de l'initiation de la transcription contrôlée en amont par des éléments de régulation sur lesquels viennent se fixer un jeu précis de facteurs de transcription.

Il aurait été intéressant de pouvoir prédire la position des sites d'initiation de la transcription dans les promoteurs pour positionner le promoteur proximal et la région 5'UTR. Il existe aujourd'hui différents programmes permettant de prédire les sites d'initiation de la transcription par l'analyse de la position des différents éléments constitutifs les uns par rapport aux autres. Malheureusement ces différents programmes ne sont pas du tout adaptés pour un génome aussi particulier que celui de *P. falciparum*. De plus, au jour d'aujourd'hui, il existe trop peu de données biologiques caractérisant les promoteurs proximaux de *P. falciparum* pour élaborer un modèle de prédiction des sites d'initiation de la transcription adapté au génome du parasite.

Comme deux TBP ont été annotées et étudiées dans le parasite en 1993 et 1994 [165, 256] et que nous avons prédit la présence d'un facteur NF-Y se fixant à la boite CCAAT dans *P. falciparum* (p. 147), il était intéressant d'analyser les séquences correspondant à ces deux types de facteurs dans les régions intergéniques de *Plasmodium*. La richesse en A+T laissait présager la présence de nombreuses boîtes TATA. En effet, les boîtes TATA sont présentes dans 99% des régions intergéniques du parasite à une fréquence moyenne de 10-11 boîtes pour 1 000 pb. Chez les eucaryotes, la boîte TATA est souvent précédée d'une boîte GC, sur laquelle se fixe la protéine Sp1, et d'une boîte CCAAT, site de fixation du facteur hétérotrimérique NF-Y entre autres. Chez *Plasmodium*, la boîte GC est inexistante si on utilise

pour la recherche le consensus eucaryote GGGGCGGGG[ct], ce qui n'est pas étonnant quand on sait que les séquences sont composées à 90% de A+T. Quant à la boîte CCAAT, elle n'existe que dans 5% des régions intergéniques, si on utilise, encore une fois, le consensus eucaryote [agc][agt]CCAAT[cg][ag] pour la recherche.

Parmi les promoteurs possédant une boîte TATA et une boîte CCAAT, seules 121 séquences présentent le schéma habituel des promoteurs proximaux des eucaryotes, à savoir une boîte TATA précédée d'une boîte CCAAT à 100 nucléotides maximum. Cependant les modules sont positionnés « au hasard » dans les régions intergéniques (données non montrées) : il semble donc impossible d'utiliser ces modules pour tenter de définir la zone du promoteur proximal. De plus, parmi ces 121 promoteurs, on ne trouve aucun des promoteurs pour lesquels le site d'initiation de la transcription a été identifié par diverses techniques biologiques (Tableau 4).

Qu'il s'agisse de la boîte TATA, de la boîte CCAAT ou de l'association des deux boîtes, il ne semble pas exister de différences entre les régions intergéniques vraies et des séquences aléatoires ayant la même composition en bases. Il est donc difficile de faire la différence entre une vraie boîte (CCAAT ou TATA) fonctionnelle et une conséquence du biais de composition. Ceci est renforcé par le fait que ces éléments et modules de régulation sont représentés différemment dans les séquences codantes qui ont u pourcentage en A+T inférieur : il y a moins de boîtes TATA, plus de boîtes CCAAT et de modules CCAAT-TATA.

Parmi tous les facteurs impliqués dans la régulation transcriptionnelle que nous avons annotés, PfMyb1 est le seul des facteurs présentés dans ce mémoire qui se fixe sur un élément de régulation précis. Et c'est le facteur que nous avons le plus étudié dans notre groupe. Comme ses homologues eucaryotes, PfMyb1 reconnaît et se fixe sur une séquence d'ADN correspondant au consensus [ct]AAC[acgt]G[act][act]. L'élément de régulation de type *myb* est présent dans la moitié des séquences intergéniques à une fréquence moyenne de moins d'un site pour 1 000 pb, ce qui est nettement inférieur à ce que l'on rencontre dans des séquences aléatoires riches en A+T et les séquences codantes. La présence de cet élément de régulation ne semble donc pas être le fruit du hasard et pourrait avoir une signification biologique.

En plus des éléments de régulation connus chez les eucaryotes, il semble logique de penser que *P. falciparum* a développé des éléments de régulation spécifiques adaptés au biais de composition. Deux groupes de promoteurs ont été utilisés pour essayer de mettre en évidence des éléments de régulation spécifiques de *Plasmodium*.

Le premier groupe est composé de 37 promoteurs gouvernant des gènes présentant un même profil d'expression, celui de *pfmyb1* (Tableau 7) : ils devraient logiquement subir la même régulation. Dans ce premier groupe, deux motifs ont été mis en évidence (Figure 26a). Ces deux motifs semblent être spécifiques de *P. falciparum* car ils sont quasiment inexistants dans des séquences aléatoires ayant la même composition en A+T que les régions intergéniques. De plus, même si ces motifs sont aussi présents dans les séquences codantes, leur représentation est différente : la moyenne est de 1 motif pour 1 000 pb dans les régions intergéniques et de 0,75 motif pour 1 000 pb dans les séquences codantes.

Le deuxième groupe est composé de promoteurs dont les gènes ont un profil d'expression altéré lorsque l'on inhibe l'ARNm *pfmyb1* et donc la production du facteur de transcription PfMyb1 (Tableau 8) : on peut donc penser que ces promoteurs sont notamment sous le contrôle, direct ou indirect, de la protéine PfMyb1. En effet, des expériences d'immunoprécipitation de la chromatine (voir Figure 7, p. 36 de l'Article 4) ont montré qu'il existait une interaction *ex vivo* entre PfMyb1 et les promoteurs de six des gènes dont le profil d'expression est altéré : MAL13P1.279, PFL1285c, PF14_0224, MAL6P1.248, MAL6P12249 et PFI1105w (alors que ce dernier ne semble pas posséder d'élément de régulation de type *myb*). L'interaction n'a pas pu être démontrée avec les deux autres promoteurs (PFL1885c et PFL2345c) et cela peut s'expliquer par les limites de la technique d'immunoprécipitation de la chromatine : soit l'interaction entre PfMyb1 et son site de liaison était trop faible pour être détectée, soit l'élément de régulation de type *myb* du promoteur était inaccessible dans la chromatine. Mais il se peut aussi que ces gènes soient sous le contrôle d'un facteur de transcription, lui-même sous le contrôle de PfMyb1 mais dont le gène n'est pas présent sur la puce à ADN thématique utilisée lors de cette expérience.

Quatre motifs ont été mis en évidence dans ces promoteurs : les deux premiers sont partagés par tous les promoteurs, les deux autres sont présents dans au moins la moitié des promoteurs. Il existe une différence de représentation entre ces quatre motifs : les motifs 1 et

2 sont très fortement représentés par rapport aux motifs 3 et 4 (3,95 et 4,37 motifs pour 1 000 pb contre 0,84 et 0,46 motifs pour 1 000 pb). Cependant, chacun de ces motifs semble avoir une signification biologique car ils sont surreprésentés par rapport à ce que l'on rencontre dans des séquences aléatoires, les motifs 3 et 4 n'existant pas du tout dans les séquences aléatoires. De plus, ces motifs ne sont pas représentés de la même manière dans les séquences codantes à cause de la différence de composition en A+T qui existe entre les séquences codantes et les séquences non codantes.

Le motif 2 identifié dans le premier groupe (Figure 26a) et le motif 3 identifié dans le deuxième groupe (Figure 27a) sont assez similaires car composés d'une suite de quatre guanines et peuvent en fait correspondre au même motif. Comme ce motif a été identifié dans les deux groupes de promoteurs ainsi que dans des promoteurs pris au hasard dans le génome, peut-être s'agit-il d'un élément de régulation constitutif qui remplacerait, chez *Plasmodium*, la boîte GC des autres eucaryotes ? Un motif hybride GGGG[agt][agt] a été recherché dans tous les promoteurs de *P. falciparum* : 6 456 motifs ont été identifiés dans 2613 séquences soit ~50% des séquences. Ce motif est quasiment inexistant dans les séquences aléatoires (507 motifs répartis dans 378 séquences) alors qu'il est très représenté dans les séquences codantes (9 626 motifs répartis dans 3 067 séquences).

Pour savoir si les motifs identifiés comme spécifiques de *P. falciparum* sont fonctionnels, c'est-à-dire s'ils permettent à un facteur de transcription de se fixer pour participer à la modulation de la transcription, il est nécessaire de valider les études *in silico* par des expériences biologiques.

Tout d'abord, il faudra commencer par faire des retardements sur gel pour savoir si une protéine ou un complexe protéique peut se fixer sur un oligonucléotide correspondant à l'élément de régulation d'intérêt. S'il y a interaction entre l'oligonucléotide et une protéine ou un complexe protéique, il sera alors indispensable d'identifier le facteur de transcription ou le complexe protéique grâce à une colonne d'affinité et de la spectrométrie de masse.

La fonctionnalité de l'élément de régulation sera ensuite étudiée grâce à la transfection d'un vecteur dans les parasites. Le vecteur sera constitué d'un gène rapporteur dont l'expression sera gouvernée par tout ou une partie du promoteur. Il sera alors nécessaire de

construire toute une série de vecteurs dans lesquels l'élément de régulation potentiel aura été muté, délété partiellement ou totalement ou encore déplacé par rapport au gène rapporteur. La comparaison du niveau d'expression du gène rapporteur dans ces différents vecteurs permettra de définir la ou les séquences et les bases directement impliquées dans le régulation et un éventuel effet de position.

L'identification d'éléments de régulation spécifiques de *P. falciparum* permettra ainsi d'identifier des facteurs de transcription spécifiques du parasite qui ne sont pas identifiables par homologie de séquences puisqu'ils n'existent pas chez les autres eucaryotes.

LES FACTEURS DE LA REGULATION TRANSCRIPTIONNELLE

Différents facteurs de transcription ont été recherchés dans le génome de *Plasmodium falciparum*, avant que celui-ci soit entièrement séquencé, c'est-à-dire avant octobre 2002.

Les facteurs recherchés ont été choisis dans la liste des facteurs donnée par la base de données TRANSFAC® [255, 404]. Notre choix s'est porté principalement sur des facteurs connus pour être impliqués dans la régulation du cycle cellulaire, du développement et de la différenciation. Néanmoins, d'autres facteurs ont été recherchés car ils étaient des coups de cœur, c'est-à-dire des facteurs sur lesquels nous avions travaillé avant ce projet ou parce qu'ils interagissaient avec un des facteurs de notre intérêt. Parmi tous les facteurs recherchés, certains ont donné de très bons résultats, alors que d'autres n'ont à ce jour toujours pas été identifiés dans le génome de *Plasmodium falciparum* avec les programmes utilisés à ce jour.

Les facteurs seront présentés non pas en fonction de la chronologie de leur annotation, ni même en fonction de la quantité des résultats obtenus mais plutôt dans l'ordre de leur intervention dans la transcription. Tout d'abord les facteurs qui jouent un rôle dans la décondensation de l'ADN permettant à la machinerie basale de transcription de se fixer juste en amont du site d'initiation de la transcription ; ensuite viendront les facteurs qui se fixent à certains éléments du promoteur proximal ; enfin, je présenterai les facteurs qui se fixent sur des séquences *cis*-régulatrices plus éloignées et qui interagissent avec la machinerie basale de transcription pour moduler, positivement ou négativement, le niveau d'expression des transcrits.

I - Facteurs de remodelage de la famille HMG

La structure de l'ADN (Figure 4) est une structure dynamique qui passe d'un état condensé à un état décondensé permettant aussi bien la transcription d'un gène que la réplication ou la réparation de l'ADN. L'état condensé, dû aux protéines histones ainsi qu'à certaines protéines non-histones, limite l'accès des facteurs de transcription à leurs séquences cibles. Il est donc indispensable que l'ADN se décondense localement, de façon à ce que la machinerie basale de transcription ainsi que les facteurs de transcription puissent se fixer à l'ADN.

Il est maintenant reconnu que certaines protéines agissent comme des facteurs de remodelage de la chromatine. Parmi ces facteurs, on trouve les histones elles-mêmes qui, par

un jeu d'acétylation-désacétylation, présentent une plus ou moins grande affinité pour la chromatine (voir p. 56). Il existe d'autres protéines de remodelage de la fibre chromatinienne qui sont des protéines chromosomiques non-histones : ce sont les petites protéines HMG (pour High Mobility Group).

I.1 - La nomenclature des protéines HMG a récemment été révisée

Une recherche bibliographique nous a appris que la nomenclature des protéines HMG avait été révisée en 2001 [49], révision qui semblait nécessaire. En effet, les protéines HMG ont été découvertes dans les cellules de mammifères il y a plus de 30 ans et appelées ainsi par rapport à leur vitesse de migration dans les gels de polyacrylamide. Des études successives ont montré que le motif fonctionnel caractéristique des protéines HMG canoniques originales se retrouvait dans des protéines nucléaires de nombreux organismes. Une façon systématique de nommer ce groupe de protéines nucléaires n'ayant alors pas encore été mise au point, le symbole d'origine 'HMG' fut utilisé pour identifier de nombreuses protéines qui n'avaient rien à voir avec les protéines HMG nucléaires.

La nomenclature des protéines HMG nucléaires a donc été révisée pour (i) faciliter les interactions entre les laboratoires, (ii) accélérer les recherches dans la littérature et (iii) éviter les confusions dues aux similitudes de noms. Les protéines HMG ont alors été divisées en trois superfamilles, chaque superfamille ayant un motif fonctionnel caractéristique :

- les protéines HMGB, dont le motif fonctionnel est appelé 'HMG-box' (anciennement HMG-1 & HMG-2),
- les protéines HMGN, dont le motif fonctionnel est appelé 'nucleosomal binding domain' (anciennement HMG-14 & HMG-17),
- les protéines HMGA, dont le motif fonctionnel est appelé 'AT-hook' (anciennement HMG-I, HMG-Y & HMG-C).

Les protéines contenant un de ces motifs fonctionnels dans leur séquence sont maintenant considérées comme des « protéines à motif HMG ».

I.2 - *Plasmodium falciparum* possède quatre protéines HMGB

Lorsque je suis arrivée au laboratoire, un premier facteur à motif HMG avait déjà été identifié sur le chromosome 12 de *P. falciparum*. Cette protéine, appelée alors HMG1, a permis l'identification par la suite de deux autres protéines : HMG2 et HMG3, respectivement sur les chromosomes 8 et 12. Puis HMG2 a permis l'identification d'une 4ème protéine sur le chromosome 13, évidemment appelée HMG4. Avec la révision de la nomenclature, nous avons pu tout de suite classer ces 4 protéines dans la superfamille des HMGB, car toutes contenaient, d'après MotifScan, un domaine 'HMG-box' (Pfam : PF00505) voire deux dans le cas de HMG3. Ces protéines ont donc été renommées PfHMGB1, PfHMGB2, PfHMGB3 et PfHMGB4 (Tableau 9).

Lorsqu'en octobre 2002, le séquençage du génome nucléaire du clone 3D7 a été terminé, nous avons pu vérifier nos résultats. Pour PfHMGB1, PfHMGB2 et PfHMGB4, aucun problème ne s'est posé. En revanche, pour PfHMGB3, nous avions identifié une phase ouverte de lecture de 5 682 nucléotides sans intron codant une protéine de 1 893 acides aminés, alors que le consortium à la base du séquençage avait identifié une séquence codante de 6 855 nucléotides, composée de 4 exons et codant une protéine de 2 284 acides aminés. Pour le moment, aucune expérience biologique n'a été effectuée nous permettant de savoir quelle annotation est correcte. Nous avons donc décidé de nous baser pour la suite sur l'annotation faite par le consortium, car elle repose sur la combinaison de plusieurs outils de prédiction de gènes.

Tableau 9. Quatre protéines HMG annotées dans le génome de *P. falciparum*.

Protéine	N° d'accession (PlasmoDB)	Taille	Localisation 'HMG-box'	Autres domaines	Localisation cellulaire
PfHMGB1	PFL0145c chr. 12	294 nt 97 aa	21 - 91	-	nucléaire (96%)
PfHMGB2	MAL8P1.72 chr. 8	300 nt 99 aa	24 - 94	-	nucléaire (96%)
PfHMGB3	PFL0290w chr. 12	6 855 nt 2 284 aa	820 - 894 895 - 964	'Myb_DNA-binding' 2 140 - 2 185	-
PfHMGB4	MAL13P1.290 chr. 13	483 nt 160 aa	6 - 80	-	nucléaire (60%)

Dans chacune de ces protéines, nous avons recherché des domaines autres que le domaine 'HMG-box' (Tableau 9) : seule la protéine PfHMGB3, la plus longue des protéines,

présente, en plus, le domaine de liaison à l'ADN caractéristique de la famille Myb (Pfam : PF00249). Mais même si aucune de ces quatre séquences ne contient de signal de localisation nucléaire, le programme PSORT a prédit que les protéines PfHMGB1, PfHMGB2 et PfHMGB4 devraient être des protéines nucléaires. Malheureusement, le programme ne fonctionne pas pour les séquences trop grandes et donc n'a donné aucun résultat pour PfHMGB3.

I.3 - Les facteurs PfHMGB sont des facteurs architecturaux

Il existe dans la superfamille HMGB (i) des **facteurs de transcription** classiques qui se fixent à l'ADN en reconnaissant une séquence précise, tels que les facteurs de transcription des cellules T TCF1 et LEF1, les protéines de typage sexuel MATa et MATα de diverses levures, le facteur SRY du déterminisme du sexe chez les mammifères ainsi que les nombreuses protéines SOX, et (ii) des protéines non-histones ou protéines HMGB classiques qui jouent probablement un rôle fonctionnel et structural dans la chromatine [35, 53, 227, 278, 289, 343, 415] et donc lors de la transcription ou la réplication de l'ADN [318]. Ces facteurs se fixent à l'ADN au niveau du petit sillon, sont capables de se fixer à de l'ADN simple ou double brin ainsi qu'à des structures d'ADN non conventionnelles comme de l'ADN modifié par le *cis*-platine [32, 110, 305], de l'ADN cruciforme [30] ou encore des jonctions entre ADN de forme B et de forme Z [158] (Figure 6c). Ces facteurs sont capables de courber l'ADN ; cette courbure semblerait faciliter la formation de complexes nucléoprotéiques importants ce qui suggère que les protéines auraient un rôle architectural dans l'assemblage de ces complexes. C'est pourquoi ils sont appelés « **facteurs architecturaux**». D'autres membres de cette superfamille tels que le facteur de transcription mitochondrial mtTF1 (mitochondrial transcription factor 1), le facteur de transcription nucléolaire UBF (upstream binding factor), la protéine SSRP1 (structure specific recognition protein) et les protéines non-histones nucléaires de levure NHP (non-histone protein) 6A & 6B présentent aussi une spécificité de séquence très faible et sont capables de reconnaître des motifs structuraux dans l'ADN.

Il y a 10 ans, on pensait que les protéines avec un seul domaine 'HMG-box' étaient des facteurs de transcription et les protéines avec plusieurs domaines 'HMG-box' des facteurs architecturaux [151]. Aujourd'hui, on sait que cette ségrégation n'est pas correcte. En effet, le

nombre de domaines 'HMG-box' chez les facteurs architecturaux est variable d'une protéine à l'autre. Cela peut aller d'un seul domaine, comme, par exemple, pour les protéines HMG-D [60, 388] et HMG-Z [275] de la drosophile ou pour les protéines NHP6A et NHP6B de *Saccharomyces cerevisiae* [211], jusqu'à cinq domaines dans le cas des protéines UBF de souris [166] ou de xénope [259], bien que la grande majorité des protéines HMGB comportent deux domaines 'HMG-box' en tandem appelés boîtes A et B. La superfamille HMGB est donc caractérisée par une extrême diversité autant fonctionnelle que structurale.

Stéphan Soullier et ses collaborateurs ont fait une analyse phylogénétique [353] montrant que les facteurs de transcription et les facteurs architecturaux peuvent être très clairement séparés sur un arbre et que cette séparation a eu lieu bien avant la divergence entre les levures et le règne animal. En effet, la superfamille HMGB est très ancienne et on trouve ses membres chez les animaux, les plantes, les levures ou les eucaryotes unicellulaires comme *Trypanosoma* [106]. Un membre de cette famille a même été transduit chez le virus *Chilo iridescens* infectant les invertébrés [337].

Je me suis donc basée sur ce travail pour faire moi-même une analyse phylogénétique de façon à savoir à quelle catégorie appartenaient PfHMGB1, PfHMGB2, PfHMGB3 et PfHMGB4. La Figure 28 représente l'arbre phylogénétique dans son ensemble.

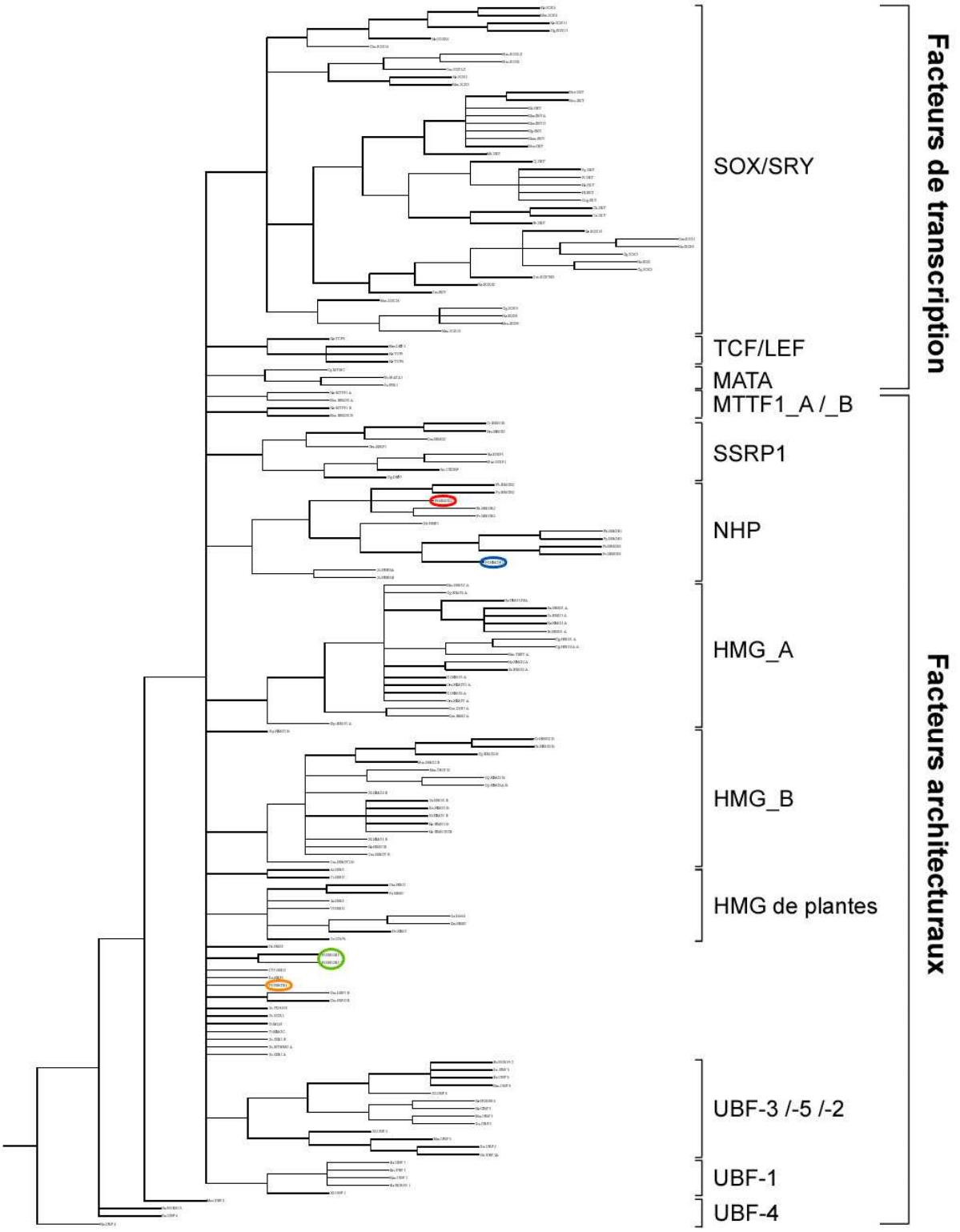


Figure 28. Arbre phylogénétique non enraciné de 159 domaines 'HMG-box' (UPGMA).
L'arbre se décompose en deux catégories : les facteurs de transcription et les facteurs architecturaux, chacune de ces catégories se divisant en plusieurs sous-familles (voir texte). Les branches apparaissant dans moins de 600 arbres sur 1000 ont été fusionnées. Sont indiqués les domaines 'HMG-box' de PfHMGB1 (en bleu), PfHMGB2 (en rouge), PfHMGB3 (en vert) et PfHMGB4 (en orange).

On retrouve la première catégorie de facteurs, à savoir les facteurs de transcription ; cette catégorie, très bien individualisée dans l'arbre complet, peut être divisée en 3 sous-familles : les protéines de typage sexuel de levures (MATA), les facteurs de transcription des cellules T (TCF/LEF) ainsi que les protéines de type SOX et SRY (SOX/SRY). La deuxième catégorie de facteurs, les facteurs architecturaux, se divise en un nombre de sous-familles beaucoup plus important :

- les boîtes A et B des protéines HMGB de vertébré (HMG_A, HMG_B) et des protéines mtTF1 (MTTF1_A et MTTF1_B),
- les domaines 'HMG-box' de plantes parmi lesquels les domaines 'HMG-box' d'*Arabidopsis thaliana* et *Catharantus roseus* se distinguent des autres plantes (HMG de plantes), des protéines SSRP1, des protéines NHP de *Saccharomyces cerevisiae* et *Babesia bovis*,
- les 5 domaines 'HMG-box' des protéines UBF (UBF-1, -2, -3, -4 et -5),

soit 12 sous-familles au total. Néanmoins, dans chacune des deux catégories, il existe des séquences qui ne se regroupent dans aucune des sous-familles décrites.

Plusieurs données sont à noter à propos de cette phylogénie et notamment au niveau des facteurs architecturaux :

- les domaines 'HMG-box' présents en plusieurs exemplaires dans une même protéine (comme les protéines HMG1 et HMG2 de vertébrés, les protéines MTTF1 ou UBF) semblent provenir d'une très ancienne duplication. En effet, les protéines HMGB de plantes ne présentent qu'un seul domaine alors que celles des métazoaires en présentent deux ;
- le gène codant une protéine HMGB classique n'existe qu'en une seule copie chez la drosophile et l'oursin, alors que chez les vertébrés, ce gène a été dupliqué une fois pour donner naissance aux gènes paralogues *hmg1* et *hmg2* (chez l'homme, les gènes *hmg1* et *hmg2* codent des protéines identiques à 80%).

Maintenant que nous avons vu l'arbre phylogénétique dans son ensemble, regardons plus précisément les protéines plasmodiales. Les facteurs sont tous les quatre localisés dans la catégorie des facteurs architecturaux. Néanmoins les domaines 'HMG-box' de PfHMGB3

et PfHMGB4 ne se regroupent dans aucune des sous-familles de facteurs architecturaux. Pour la suite du travail, je me suis intéressée plus particulièrement aux facteurs PfHMGB1 et PfHMGB2.

Les facteurs HMGB1 et HMGB2 de toutes les espèces de *Plasmodium* forment deux sous-arbres distincts. Alors que, dans les deux cas, les séquences des parasites murins *P. berghei* et *P. yoelii* se regroupent ensemble, les séquences de *P. vivax*, parasite infectant l'homme, ne se regroupent pas avec les séquences de *P. falciparum* mais avec les séquences du parasite simien *P. knowlesi*. Néanmoins, toutes les séquences plasmodiales se regroupent avec la séquence NHP1 de *Babesia bovis*, un parasite des globules rouges humains transmis par les tiques et qui possède aussi un génome très riche en A+T, ainsi qu'avec les séquences NHP6A et NHP6B de *Saccharomyces cerevisiae*, la levure de boulanger (Figure 29). Une remarque est cependant à faire ici : la phylogénie présentée ici a été faite avec la méthode UPGMA alors que la phylogénie présentée dans l'article *High-Mobility-Group box nuclear factors of Plasmodium falciparum* a été faite avec la méthode Neighbor-joining. Les topologies générales des deux arbres obtenus sont très similaires mais dans l'arbre obtenu avec la méthode Neighbor-joining, les domaines 'HMG-box' des protéines NHP6 de levure ne sont pas regroupés avec les domaines parasitaires et forment un groupe à part.

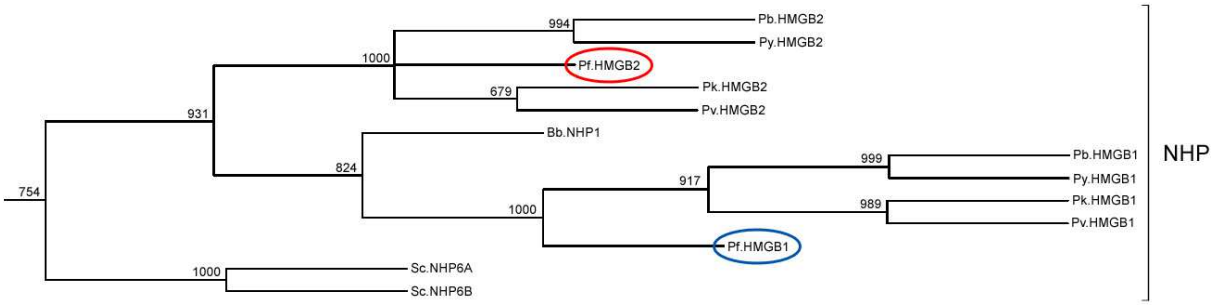


Figure 29. Détail de l'arbre phylogénétique des facteurs de la famille HMGB.
Les noms des facteurs sont codés de la manière suivante : « abréviation_organisme.nom_facteur » (Bb, *Babesia bovis* ; Pb, *Plasmodium berghei* ; Pf, *Plasmodium falciparum* ; Pk, *Plasmodium knowlesi* ; Pv, *Plasmodium vivax* ; Py, *Plasmodium yoelii* ; Sc, *Saccharomyces cerevisiae*). A chaque nœud sont indiquées les valeurs de bootstrap (valeur maximale : 1000).

Peu de choses sont connues à propos de la protéine NHP1 [75] : c'est une protéine de 97 acides aminés, comme PfHMGB2, qui présente une identité de séquence de 45% avec NHP6A, ce pourcentage augmentant à 57% quand on ne considère que les domaines 'HMG-

box'. En revanche, les protéines NHP6 de *Saccharomyces cerevisiae* ont été beaucoup plus étudiées depuis de nombreuses années.

En 1988, David Kolodrubetz identifie NHP6, une protéine non-histone similaire à la protéine HMG1 de thymus de veau [212]. Dès 1990, il se rend compte qu'il existe en fait deux gènes *nhp6* sans intron, qui sont tous les deux transcrits et codent des protéines très similaires : NHP6B possède six acides aminés supplémentaires situés en N-terminal par rapport à NHP6A mais les deux protéines présentent 87% d'identité sur le reste de leurs séquences [211] ; néanmoins, ces protéines sont plus petites que les protéines HMG des eucaryotes supérieurs précédemment identifiées. Quelques années plus tard, Paull & Johnson montrent que ces deux protéines sont des protéines architecturales car elles sont capables de se lier à l'ADN, de le courber voire de créer un superenroulement, et semblent former des complexes ADN-protéine très stables [297].

Pour savoir si les protéines PfHMGB1 et PfHMGB2 présentent les mêmes caractéristiques que les protéines de différents organismes, dont les levures, les séquences ont été comparées grâce un alignement multiple des domaines 'HMG-box'. Il en ressort que les domaines 'HMG-box' sont assez bien conservés au cours de l'évolution (Tableau 10 et Figure 30), l'alignement ayant été fait avec des séquences de parasites, de levure, d'insecte, de plantes et de mammifères. De nombreux résidus sont retrouvés dans la majorité des séquences et s'ils ne sont pas identiques, ils sont fortement similaires. On remarque aussi que les protéines HMGB1 et HMGB2 sont extrêmement bien conservées d'une espèce plasmodiale à l'autre.

Tableau 10. Pourcentage d'identité entre les domaines 'HMG-box' de PfHMGB1 & PfHMGB2 et les domaines 'HMG-box' de différentes protéines.

	PfHMGB1		PfHMGB2	
Dm.HMG-D	35,21 %	(54,93 %)	35,21 %	(57,75 %)
Rn.HMG1_A	36,49 %	(52,71 %)	29,73 %	(55,41 %)
Hs.HMG1_A	36,49 %	(52,71 %)	29,73 %	(55,41 %)
Hs.HMG2_A	35,14 %	(54,06 %)	31,08 %	(58,11 %)
Zm.MNB1B	44,44 %	(65,27 %)	41,67 %	(63,89 %)
Os.HMGB1	43,66 %	(64,79 %)	42,25 %	(63,38 %)
Gm.HMG1	38,03 %	(61,97 %)	40,85 %	(61,98 %)
Rn.HMG1_B	42,25 %	(61,97 %)	39,44 %	(61,98 %)
Hs.HMG1_B	42,25 %	(61,97 %)	39,44 %	(61,98 %)
Hs.HMG2_B	40,85 %	(63,39 %)	38,03 %	(61,97 %)
Cg.HMG1	42,25 %	(61,97 %)	39,44 %	(61,98 %)
Sc.NHP6A	52,11 %	(70,42 %)	51,39 %	(70,83 %)
Sc.NHP6B	50,70 %	(70,42 %)	47,22 %	(69,44 %)
Bb.NHP1	66,20 %	(91,55 %)	59,15 %	(78,87 %)

Dans chaque colonne, le premier chiffre correspond au pourcentage d'identités et entre parenthèses est indiqué le pourcentage d'identités et de similitudes fortes.

De plus, en 1999, en étudiant des facteurs de transcription et des facteurs architecturaux, Frank Murphy IV et ses collaborateurs ont pu mettre en évidence 3 déterminants caractéristiques de chacun des deux groupes de protéines [266]. Les deux premiers se situent sur la séquence : aux positions 10 et 32 selon la numérotation de la protéine HMG-D de la drosophile (positions encadrées sur la Figure 30), les facteurs de transcription possèdent une asparagine et un acide aminé hydrophile alors que les facteurs architecturaux présentent une serine et un acide aminé hydrophobe. Le troisième déterminant est la présence d'un cœur hydrophobe chez les facteurs de transcription donc celui-ci n'est identifiable que si la structure tridimensionnelle de la protéine est connue. Si on se réfère à l'alignement, on voit que toutes les protéines plasmodiales appartiennent au groupe des facteurs architecturaux, ce qui renforce les résultats obtenus avec la phylogénie.

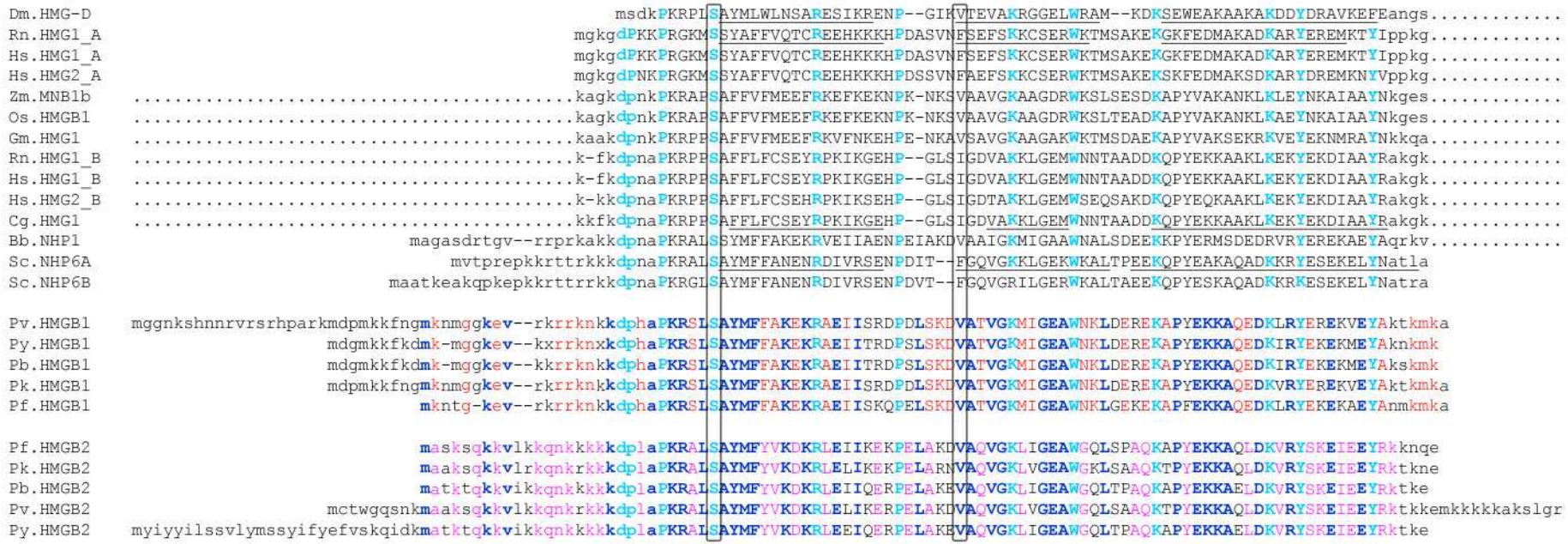


Figure 30. Alignement multiple des séquences complètes des protéines de *Plasmodium*, de *S. cerevisiae* et de *B. bovis* avec les domaines 'HMG-box' des protéines de différents organismes.

Les points et les tirets représentent respectivement les parties de séquences qui n'ont pas été inclues dans l'alignement et les insertions-délétions introduites par le programme MultAlin et nécessaires au bon alignement des séquences. En majuscule sont indiqués les domaines 'HMG-box' identifiés par le programme MotifScan. Les résidus conservés dans plus de 95% de toutes les séquences sont indiqués en bleu clair, les résidus conservés dans toutes les séquences plasmodiales en bleu foncé, les résidus conservés dans toutes les séquences HMGB1 ou HMGB2 de *Plasmodium* en rouge ou rose. Les résidus encadrés sont deux des trois déterminants mis en évidence par Frank Murphy IV et ses collaborateurs [266]. Les résidus soulignés sont ceux qui font partie des hélices α (selon les données obtenues sur PDBSum avec les fichiers 1QRV pour Dm.HMG-D, 1CKT pour Rn.HMG1_A, 1HSN pour Cg.HMG1 et 1J5N pour Sc.NHP6A). Les noms des facteurs sont codés comme pour la Figure 29. Les abréviations sont les suivantes : Bb, *Babesia bovis* ; Cg, *Cricetulus griseus* ; Dm, *Drosophila melanogaster* ; Gm, *Glycine max* ; Rn, *Rattus norvegicus* ; Os, *Oryza sativa* ; Pb, *Plasmodium berghei* ; Pf, *Plasmodium falciparum* ; Pk, *Plasmodium knowlesi* ; Pv, *Plasmodium vivax* ; Py, *Plasmodium yoelii* ; Sc, *Saccharomyces cerevisiae* ; Zm, *Zea mays*.

Lorsque l'on fait un alignement avec les séquences complètes et non plus uniquement avec les domaines 'HMG-box' (Figure 31), on se rend compte de plusieurs choses. Tout d'abord, l'unique domaine 'HMG-box' de PfHMGB1 et PfHMGB2, tout comme celui des protéines de levure, de drosophile et de plantes, est plus proche de la boîte B que de la boîte A des protéines de vertébrés possédant deux domaines 'HMG-box'. Ceci a aussi été démontré par un arbre phylogénétique généré uniquement avec les boîtes A et B des protéines de vertébrés et les domaines 'HMG-box' des protéines parasitaires. Deux groupes distincts apparaissent : un premier regroupant les boîtes B et les domaines 'HMG-box' de PfHMGB1 et PfHMGB2, un deuxième regroupant toutes les boîtes A (voir donnée additionnelle de l'article *High-Mobility-Group box nuclear factors of Plasmodium falciparum*, 2006, p.).

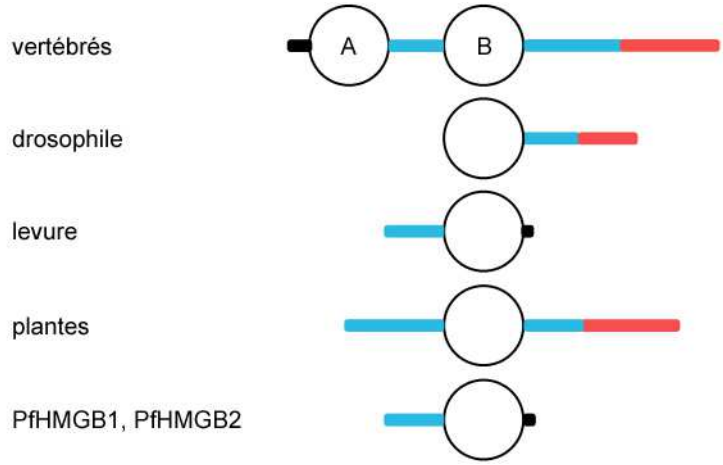


Figure 31. Représentation schématique de l'alignement multiple des séquences complètes de facteurs architecturaux.
Les cercles représentent les domaines 'HMG-box' de chaque groupe de protéines, les facteurs HMGB de vertébrés ayant le plus souvent deux boîtes nommées A et B alors que les autres n'en possèdent qu'une. En bleu sont représentés les domaines basiques des protéines, en rouge les queues acides et en noir les parties de séquences sans particularité. Cette image a été adaptée d'une figure de la revue de Thomas & Travers [374].

Ensuite, les protéines PfHMGB sont caractérisées (i) par la présence d'une région basique en N-terminal du domaine 'HMG-box' (en bleu sur la Figure 31), comme les facteurs de plantes et de levure et (ii) par l'absence de région basique en C-terminal du domaine 'HMG-box' et de queue acide (en rouge sur la Figure 31), comme les facteurs de la levure.

I.4 - PfHMGB1 & PfHMGB2 ont un domaine de liaison à l'ADN en forme de L

La structure tridimensionnelle de quelques facteurs HMGB est connue. Nous avons donc décidé de modéliser la structure des facteurs PfHMGB1 & PfHMGB2 par homologie à des structures connues, selon la méthode expliquée dans les Matériels et Méthodes (p. 89).

La méthode « tout-automatique » utilisée pour modéliser les facteurs PfHMGB1 et PfHMGB2 a désigné comme structure support la boîte B du facteur HMG1 de *Cricetulus griseus*, le hamster chinois, obtenue par spectroscopie RMN (fichier PDB : 1HSN) [317]. Le fragment de protéine de 79 acides aminés, qui est capable de fixer de l'ADN cruciforme, est constitué de 4 hélices α (appelées α1, α1', α2 et α3) qui se replient pour former un L et ne présente pas de brins β. L'échantillon utilisé pour la spectroscopie RMN contient une molécule de β-mercaptoéthanol (βME) fixée à la seule cystéine du domaine 'HMG-box'.

Une partie des facteurs PfHMGB1 et PfHMGB2 a été modélisée, respectivement de l'histidine 19 à la tyrosine 90 et de l'alanine 23 à la glutamine 98, grâce à des alignements faits entre la séquence cible et la séquence support (Figure 32).

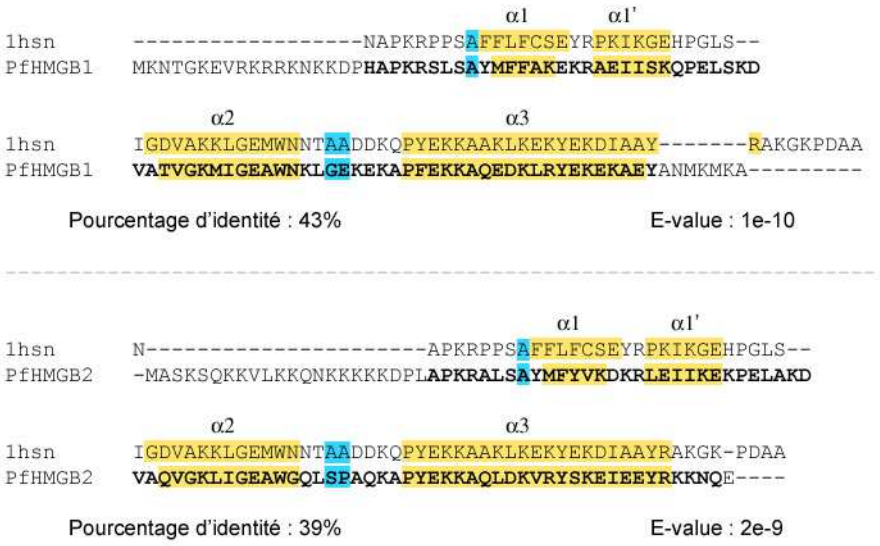


Figure 32. Alignements entre les séquences cibles PfHMGB et la séquence support 1HSN.
En gras sont indiquées les parties des séquences cibles qui ont été modélisées par homologie avec la séquence de la structure support et en jaune les acides aminés faisant partie des hélices α. α1, α1', α2 et α3 sont les noms des 4 hélices de la structure de la boîte B du facteur HMG1 de *Cricetulus griseus* (fichier PDB : 1HSN). En bleu sont indiqués les résidus qui se situent en dehors des zones autorisées du diagramme de Ramachandran.

Deux points sont à noter sur ces alignements :

- dans les deux cas, deux insertions ont été introduites dans la séquence support entre l'hélice α1' et l'hélice α2. Ces insertions auront pour effet d'augmenter la taille de la boucle située entre les deux hélices.
- les hélices des modèles ne sont pas de la même taille que les hélices de la structure support : les hélices α1, α2 et α3 du modèle de PfHMGB1 ainsi que les hélices α1 et α2 du modèle de PfHMGB2 sont plus petites d'un ou deux résidus.

Les structures modèles obtenues contiennent elles aussi 4 hélices α, repliées en forme de L (Figure 33). Normalement les domaines 'HMG-box' comportent 3 hélices α mais, ici, la molécule de βME « casse » la première hélice en hélices α1 et α1'.

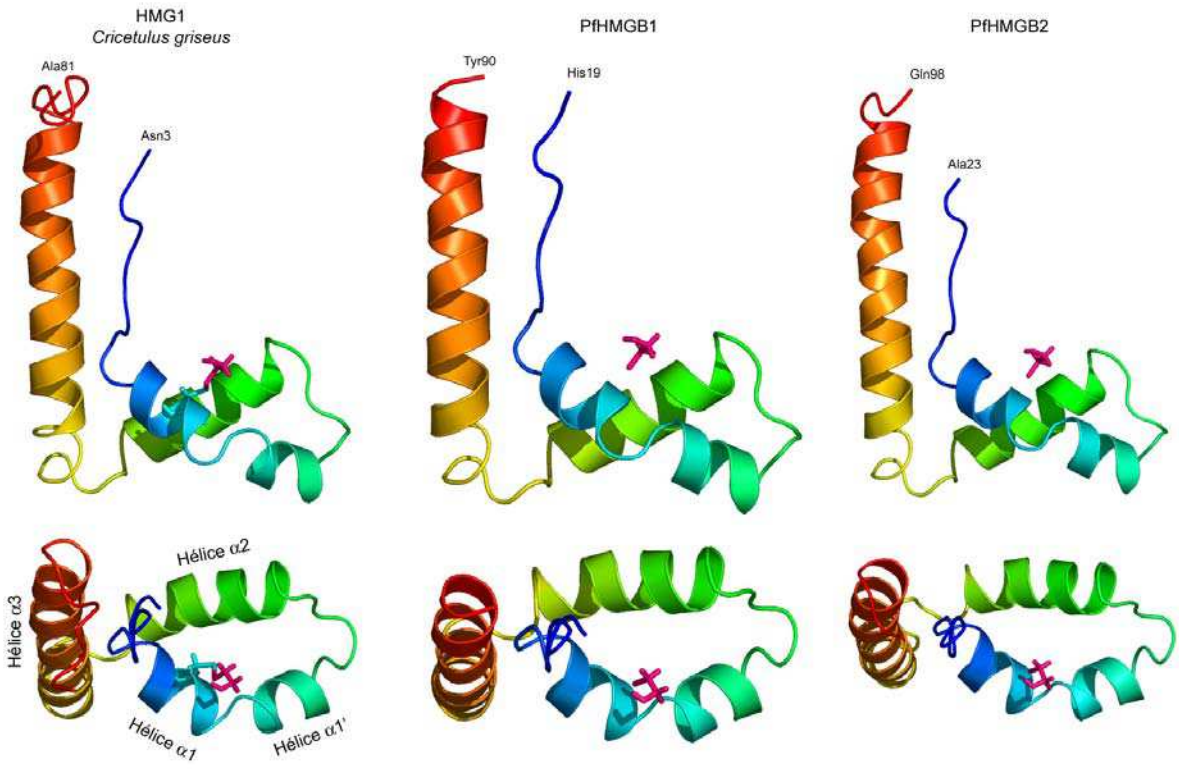


Figure 33. Modèles des facteurs PfHMGB1 & PfHMGB2 obtenus par homologie avec le facteur HMG1 du hamster chinois *Cricetulus griseus*.
Chaque structure, à savoir la structure support et les deux structures modèles, est visualisée sous deux angles différents. Les trois structures sont colorées selon le spectre de la lumière, le bleu désignant la partie N-terminale de la structure et le rouge la partie C-terminale. En rose est figuré la molécule de βME, initialement fixée à la cystéine (en cyan) située dans la première hélice de la structure du hamster chinois.

Quand les structures modèles obtenues sont comparées à la structure support, les boucles situées entre les hélices α1' et α2 apparaissent de même taille (Figure 33) : en effet, quand on regarde plus en détail les fichiers PDB générés par TITO pour les modèles de PfHMGB1 et PfHMGB2, on se rend compte que la paire d'acides aminés KD correspondant aux insertions dans la séquence support n'a pas été modélisée. Ceci est confirmé par les résultats du logiciel CE : lorsqu'il compare une structure modèle à la structure support, il trouve un RMSD de 0,0 Å et un Z-score de 5,6 pour un alignement de 70 résidus sans gap pour PfHMGB1 et un RMSD de 0,0 Å et un Z-score de 5,7 pour un alignement de 74 résidus sans gap pour PfHMGB2. Ceci signifie plusieurs choses : (i) les carbones α de la structure modèle sont très bien superposés à ceux de la structure support mais les insertions dans la boucle située entre les hélices α1' et α2 ont été éliminées dans les modèles et (ii) les séquences présentent un degré d'identité /similitude tel que les séquences peuvent être considérés comme appartenant à la même famille (d'après la valeur de Z-score supérieure à 4,5).

Certains résidus des modèles de PfHMGB1 & PfHMGB2 (surlignés en bleu sur la Figure 32) sortent des zones autorisées du diagramme de Ramachandran mais ils correspondent à des résidus de la structure support qui, eux aussi, sortaient des zones autorisées. Néanmoins, d'après Verify3D et ProSa2003, les modèles sont qualitativement moins bons que la structure support car ils présentent des scores moins « intéressants » (données non montrées).

Après la méthode « tout-automatique », une autre méthode qui consiste à tout faire, ou presque, « soi-même » a été utilisée. Les séquences les plus proches soit du domaine 'HMG-box' de PfHMGB1 ou PfHMGB2 soit de la séquence protéique de chaque facteur dans sa totalité ont été recherchées dans la base de données PDB. Toutes les requêtes ont donné les mêmes résultats (Tableau 11), mais pas forcément dans le même ordre.

Lorsque la requête a été effectuée avec la séquence du facteur PfHMGB1, les réponses sont arrivées dans l'ordre indiqué dans le Tableau 11. Lorsque cette requête a été effectuée avec la séquence de PfHMGB2, les réponses ont été obtenues dans un ordre différent, néanmoins les réponses concernant la protéine NHP6A de *S. cerevisiae* sont toujours arrivées en tête. C'est pourquoi mon choix s'est porté sur la protéine NHP6A car elle correspond à

une protéine entière, de taille similaire aux facteurs de *P. falciparum*, et ne présente pas de mutation. En plus de sa très bonne E-value, son pourcentage d'identité avec PfHMGB1 & PfHMGB2 en fait un très bon candidat pour être la structure support qui va permettre de modéliser les deux facteurs parasitaires. Parmi les trois structures de NHP6A disponibles, la structure 1J5N a été choisie car c'est la seule qui a été déterminée avec un double brin d'ADN.

Tableau 11. Structures PDB dont la séquence est homologue aux facteurs PfHMGB.

N° accession PDB / SwissProt	Description	Méthode	Remarques	Données sur les résultats de la requête BlastP			
				Longueur	Score	% identité	E-value
1LWM / P11632	*S. cerevisiae* NHP6A	RMN, 20 (1)	protéine entière	93 / 93	217 / 216	52 / 53	2e-16 / 2e-16
1J5N / P11632	***S. cerevisiae* NHP6A**	**RMN, 20 (1)**	**protéine entière, avec double hélice d'ADN**	**93 / 93**	**217 / 216**	**52 / 53**	**2e-16 / 2e-16**
1CG7 / P11632	*S. cerevisiae* NHP6A	RMN, 30 (1)	protéine entière	93 / 93	217 / 216	52 / 53	2e-16 / 2e-16
1J3D / P17741	*S. scrofa* HMGB2	RMN, 30 (1)	boîte B mutant H108Y	78 / 78	164 / 151	44 / 41	6e-11 / 1e-09
1J3C / P17741	*S. scrofa* HMGB2	RMN, 30 (1)	boîte B	79 / 79	163 / 155	44 / 55	8e-11 / 6e-10
1HMF / P63159	*R. norvegicus* HMG1	RMN, 30	boîte B	77 / 77	161 / 146	45 / 42	1e-10 / 5e-09
1HME / P63159	*R. norvegicus* HMG1	RMN, 1	boîte B modèle 24 de 1HME	77 / 77	161 / 146	45 / 42	1e-10 / 5e-09
1AAB / P07155	*R. norvegicus* HMG1	RMN, 33 (2)	boîte A mutant C22S	83 / 83	146 / 132	38 / 32	5e-09 / 1e-07
1J3X / P17741	*S. scrofa* HMGB2	RMN, 30 (1)	boîte A	77 / 77	145 / 127	38 / 33	6e-09 / 5e-07
1NHN / P07156	*C. griseus* HMG1	RMN, 41	boîte B	81 / 81	143 / 132	43 / 39	1e-08 / 1e-07
1NHM / P07156	*C. griseus* HMG1	RMN, 1	boîte B structure minimale parmi les 41 de 1NHN	81 / 81	143 / 132	43 / 39	1e-08 / 1e-07
1HSN / P07156	*C. griseus* HMG1	RMN, 49	boîte B, avec du βME	79 / 79	143 / 132	43 / 39	1e-08 / 1e-07
1HSM / P07156	*C. griseus* HMG1	RMN, 1	boîte B, avec du βME structure minimale parmi les 49 de 1HSM	79 / 79	143 / 132	43 / 39	1e-08 / 1e-07
1CKT / P07155	*R. norvegicus* HMG1	RX (2,5Å)	boîte A, avec ADN modifié par le *cis* -platine	71 / 71	129 / 116	37 / 31	3e-07 / 7e-06

La colonne 'Méthode' contient différentes informations : la technique utilisée pour obtenir la structure de la protéine (RMN ou RX pour cristallographie aux rayons X), suivie, dans le cas de la RMN, du nombre de structures dans le fichier PDB avec le n° du meilleur modèle entre parenthèses, ou dans le cas de la cristallographie, de la résolution du cristal entre parenthèses. Dans la colonne 'Remarques' sont regroupées toutes les informations à prendre en considération pour le choix de la structure la plus appropriée pour la modélisation. La colonne 'Données sur les résultats de la requête BlastP' est séparée en quatre sous-colonnes : la longueur de l'alignement entre la séquence requête et la séquence homologue, le score, le pourcentage d'identité et la E-value associés à cet alignement. Dans chaque sous-colonne, le premier chiffre correspond aux résultats obtenus avec comme requête la séquence de PfHMGB1 et le deuxième chiffre avec celle de PfHMGB2.

Pour préparer les fichiers nécessaires au programme Modeller, plusieurs alignements entre la séquence cible et la séquence support (2 à 2, multiple et structural) ont été faits puis comparés pour en faire un « alignement consensus » auquel différentes informations comme les résidus en contact avec l'ADN et/ou essentiels à l'activité biologique de la protéine, la position des structures secondaires vraies et prédites, etc. ont été intégrées (Figure 34). On

peut d'ailleurs remarquer sur cet alignement que la prédiction des structures secondaires ne correspond pas toujours à la réalité, en terme de longueur et de position.

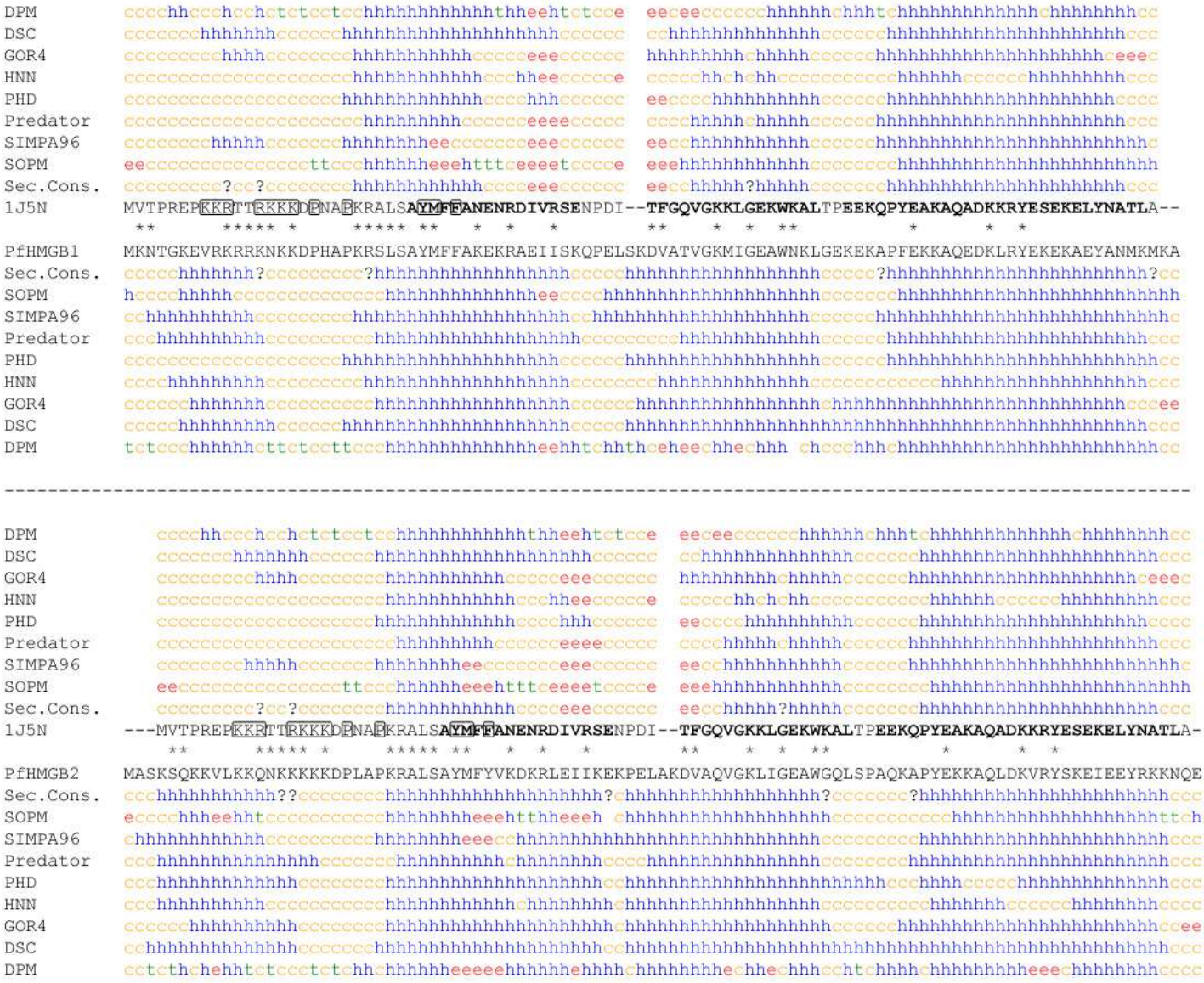


Figure 34. Alignement de PfHMGB1 & PfHMGB2 avec la séquence support NHP6A.
En noir se trouvent les séquences cibles PfHMGB1 et PfHMGB2 et la séquence de la structure support 1J5N. Les résidus encadrés sont essentiels à la stabilité du repliement de la protéine NHP6A et de sa liaison à l'ADN [412]. Les caractères gras indiquent les hélices α réelles de la structure support et les astérisques (*) les résidus de NHP6A en contact avec l'ADN (fichier PDB 1J5N). Chaque séquence a été soumise au programme CONSENSUS qui regroupe différents outils de prédiction de structures secondaires. Sont représentés : en bleu les hélices, en rouge les brins, en vert les β-turns, en jaune les coils et par un ? les états ambigus dans le consensus des prédictions de structures secondaires.

Les acides aminés importants de la structure support, c'est-à-dire les acides aminés indispensables à la stabilité du repliement de la protéine ou encore nécessaires à la fixation de la protéine sur l'ADN sont pour la plupart identiques ou similaires dans les séquences cibles. De plus, le gap de deux résidus intégré dans la séquence de levure a été décalé d'un résidu pour être placés dans une boucle et non plus dans une hélice par comparaison avec l'alignement multiple effectué au début (Figure 30).

Parmi les 100 modèles générés par Modeller pour chacune des deux protéines parasitaires, deux modèles ont été retenus pour PfHMGB1 (modèles 1.1, 1.2) et quatre pour PfHMGB2 (2.1, 2.2, 2.3, 2.4) grâce à la valeur de leur fonction objective (voir Figure 64, p. 236).

Etudions tout d'abord les résultats de la simulation pour PfHMGB1. Les deux structures modèles de PfHMGB1 présentent des valeurs de RMSD faibles lorsqu'on les compare à la structure support avec le programme CE (1,6 et 1,4 Å respectivement) et une valeur de 1,2 Å quand on les compare l'une à l'autre. Ces deux structures modèles sont donc très proches de la structure support mais surtout très proches l'une de l'autre. De plus, d'après Verify3D et ProSa2003 (voir Figure 65, p. 237), les deux modèles sont qualitativement très corrects par rapport à la structure support : ProSa2003 indique que les modèles sont quelquefois moins bons et quelquefois meilleurs que la structure support selon la zone des structures sur laquelle on se focalise alors que Verify3D montre que les deux modèles sont qualitativement meilleurs que la structure support. Les résultats de ces deux programme sont assez concordants : en effet, les scores les plus élevés de Verify3D correspondent aux scores les plus faibles de ProSa2003, et inversement. De plus, les diagrammes de Ramachandran des deux structures modèles indiquent qu'aucune ne contient de résidus situés en dehors des zones autorisées, tout comme la structure support (données non montrées)

Dans l'ensemble, les deux structures modèles se superposent assez bien à la structure support (Figure 35a). Les seules différences que l'on peut observer sont les suivantes :

- La partie N-terminale de la protéine PfHMGB1 (Figure 35b) ne présente pas de structure secondaire bien définie et peut donc adopter différentes conformations. Les 18 premiers acides aminés se replient de manière complètement différente dans les deux structures modèles et ce sans suivre le repliement de la structure support.

Néanmoins, à partir de la proline 18, les trois structures se rejoignent pour quasiment se confondre juste avant de former la première hélice α.

- Les deux structures modèles se replient aussi de manière différente en C-terminal, après la troisième hélice α. Cette partie correspond aux deux acides aminés supplémentaires de PfHMGB1 qui ne trouvent donc pas de support pour être modélisés.
- La boucle située entre les hélices α1 et α2 est différente dans les deux modèles (Figure 35c) car elle comporte la partie de la protéine correspondant aux deux insertions ajoutées dans la séquence support (Figure 34). Le programme Modeller doit donc modéliser cette partie de la protéine qui se situe entre deux structures fixes très bien conservées sans avoir de modèle à suivre. Dans le cas du modèle 1.1, le programme Modeller a choisi de conserver la longueur de l'hélice α1 et d'allonger la boucle alors que dans le modèle 1.2, l'hélice α1 a été allongée et on peut voir qu'elle est vrillée.

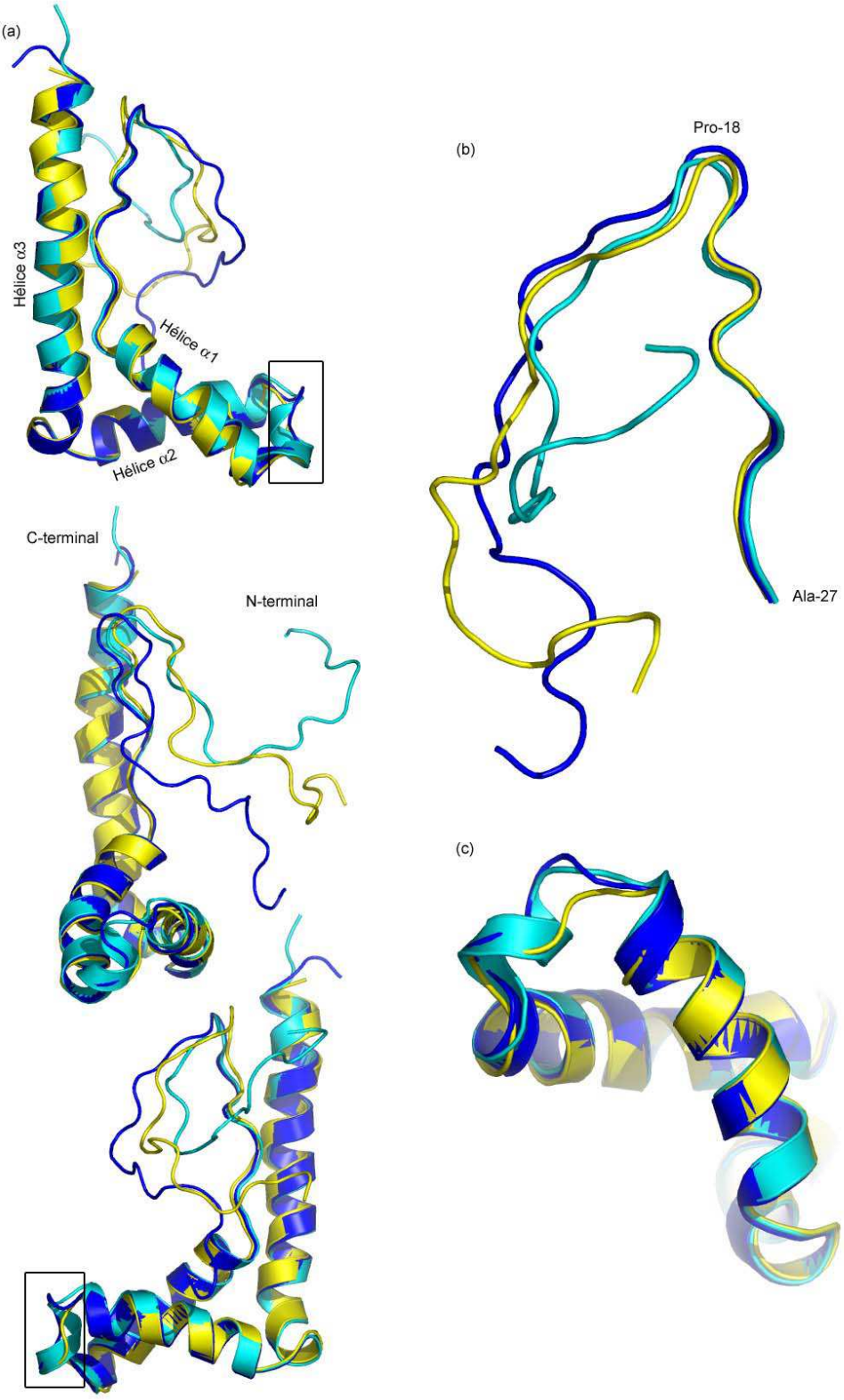


Figure 35. Les deux structures modèles de PfHMGB1 avec la structure support.
La structure support 1J5N est colorée en jaune, le modèle 1.1 en bleu et le modèle 1.2 en cyan. Les structures ont été superposées grâce à l'option 'Iterative Magic Fit' de SwissPDB Viewer. **(a)** Structures entières montrées sous trois angles différents autour d'un même axe de rotation. **(b)** Partie N-terminale avant la première hélice α de chaque structure. La numérotation est faite selon la séquence PfHMGB1. **(c)** Zoom de la boucle située entre les hélices α1 et α2 (encadré sur (a)).

En ce qui concerne PfHMGB2, lorsque les structures modèles sont comparées à la structure support avec le programme CE, trois des quatre structures modèles de PfHMGB2 présentent des valeurs de RMSD situées entre 1,0 et 1,2 Å alors que le modèle 2.3 présente une valeur plus élevée (2,2 Å). L'analyse des modèles faite par le programme Verify3D montre que les 4 modèles sont moins bons que la structure support sur la première moitié de la protéine mais meilleurs sur la deuxième partie alors que les résultats obtenus avec ProSa2003 indiquent que les structures modèles sont équivalentes à la structure support sur l'ensemble de la protéine (voir Figure 66, p. 238). De plus les diagrammes de Ramachandran des quatre structures modèles indiquent qu'aucune ne contient de résidus situés en dehors des zones autorisées (données non montrées).

Les structures modèles de PfHMGB2 se superposent assez bien à la structure support (Figure 36a) surtout au niveau des trois hélices α, mais il existe certaines différences :

- Pour la partie N-terminale de la protéine PfHMGB2, Modeller propose différentes configurations (Figure 36b). Les modèles 2.1, 2.2 et 2.4 ont des repliements différents et assez éloignés spatialement sur les 14 premiers acides aminés mais à partir de la lysine 15, les structures se rejoignent jusqu'à la sérine 29, dernier acide aminé avant la première hélice α. Quant au modèle 2.3, son repliement est complètement différent sur les 20 premiers acides aminés, ce qui expliquerait la valeur élevée de RMSD trouvée par le programme CE.
- Les quatre structures modèles ont une boucle située entre les hélice 1 et 2 plus longue que celle de la structure support (Figure 36c). Dans le cas de PfHMGB2, le programme Modeller n'a jamais allongé la première hélice mais a toujours allongé la boucle. D'ailleurs deux trajectoires ont été retenues par le programme, et non quatre, et que dans le cas des modèles 2.3 et 2.4, la deuxième hélice commence plus tôt.

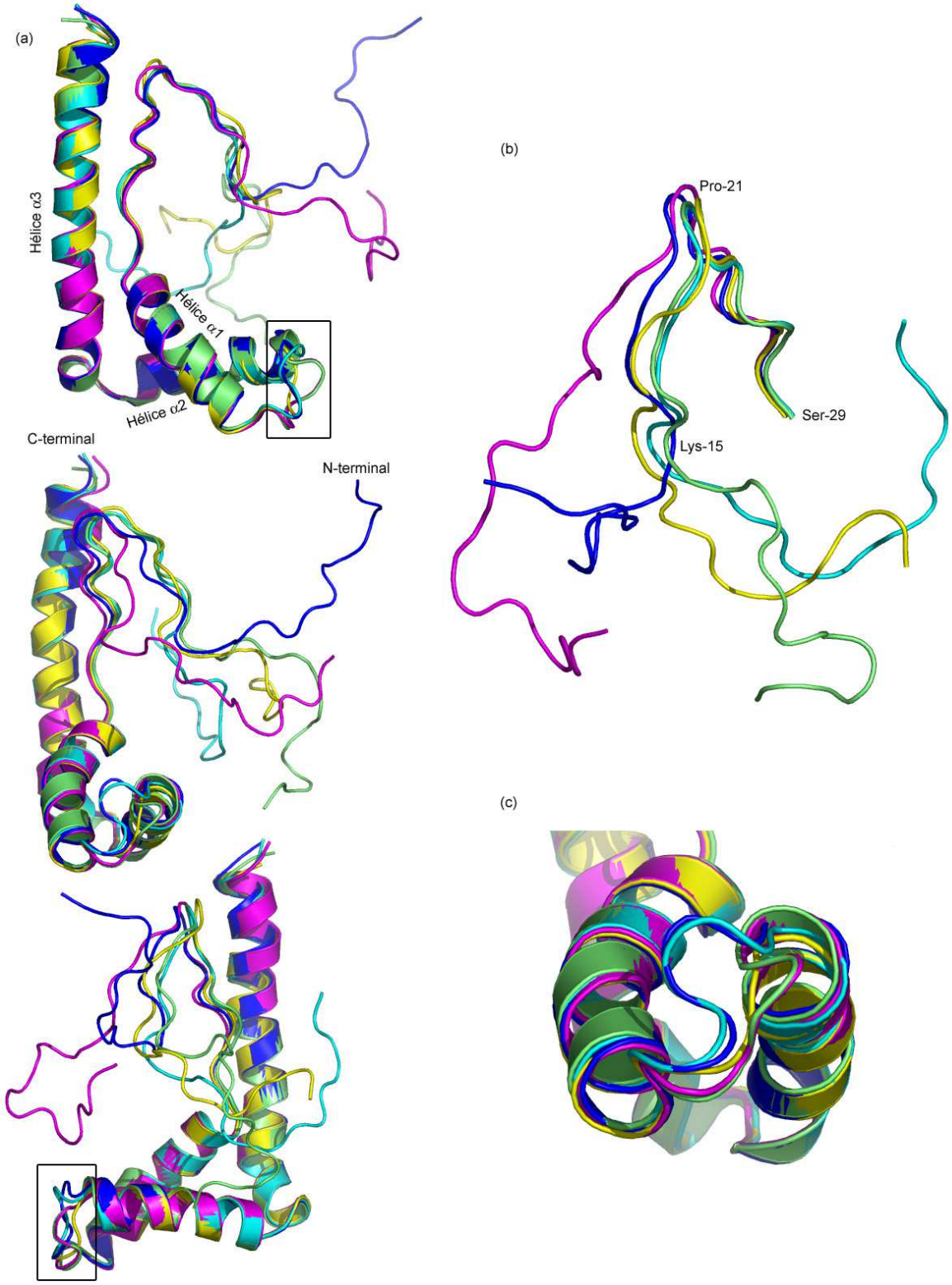


Figure 36. Les quatre structures modèles de PfHMGB2 avec la structure support.
La structure support 1J5N est colorée en jaune, le modèle 2.1 en bleu, le modèle 2.2 en cyan, le modèle 2.3 en magenta et le modèle 2.4 en vert. Les structures ont été superposées grâce à l'option 'Iterative Magic Fit' de SwissPDB Viewer. **(a)** Structures entières montrées sous trois angles différents autour d'un même axe de rotation. **(b)** Partie N-terminale avant la première hélice α de chaque structure. La numérotation est faite selon la séquence PfHMGB2. **(c)** Zoom de la boucle située entre les hélices α1 et α2 (encadré sur (a)).

Dans le fichier PDB 1J5N se trouvent les coordonnées atomiques d'un double brin d'ADN sur lequel est fixée la protéine NHP6A (Figure 37a). La suite du travail a donc consisté à voir comment les modèles pouvaient interagir avec ce double brin d'ADN.

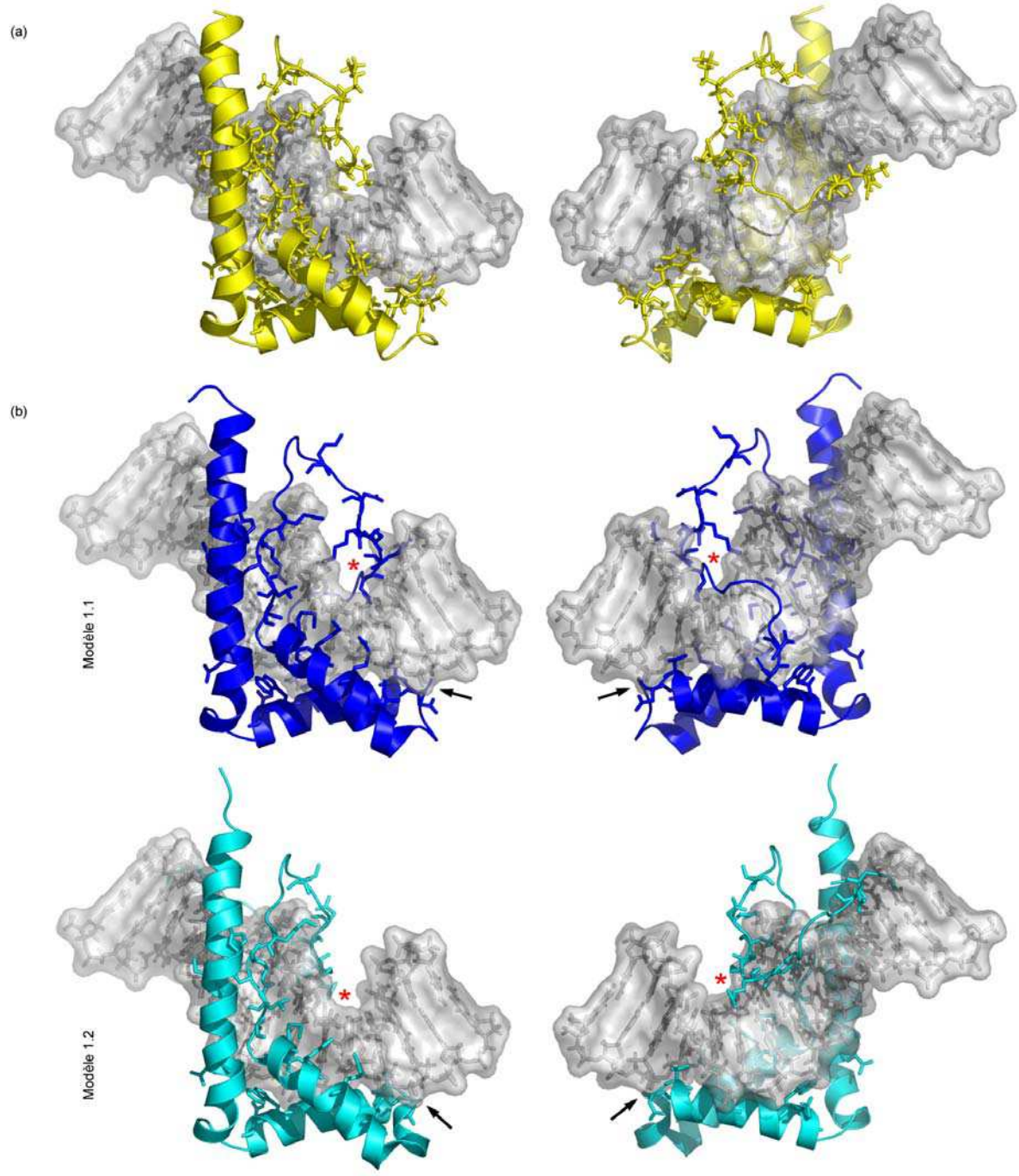
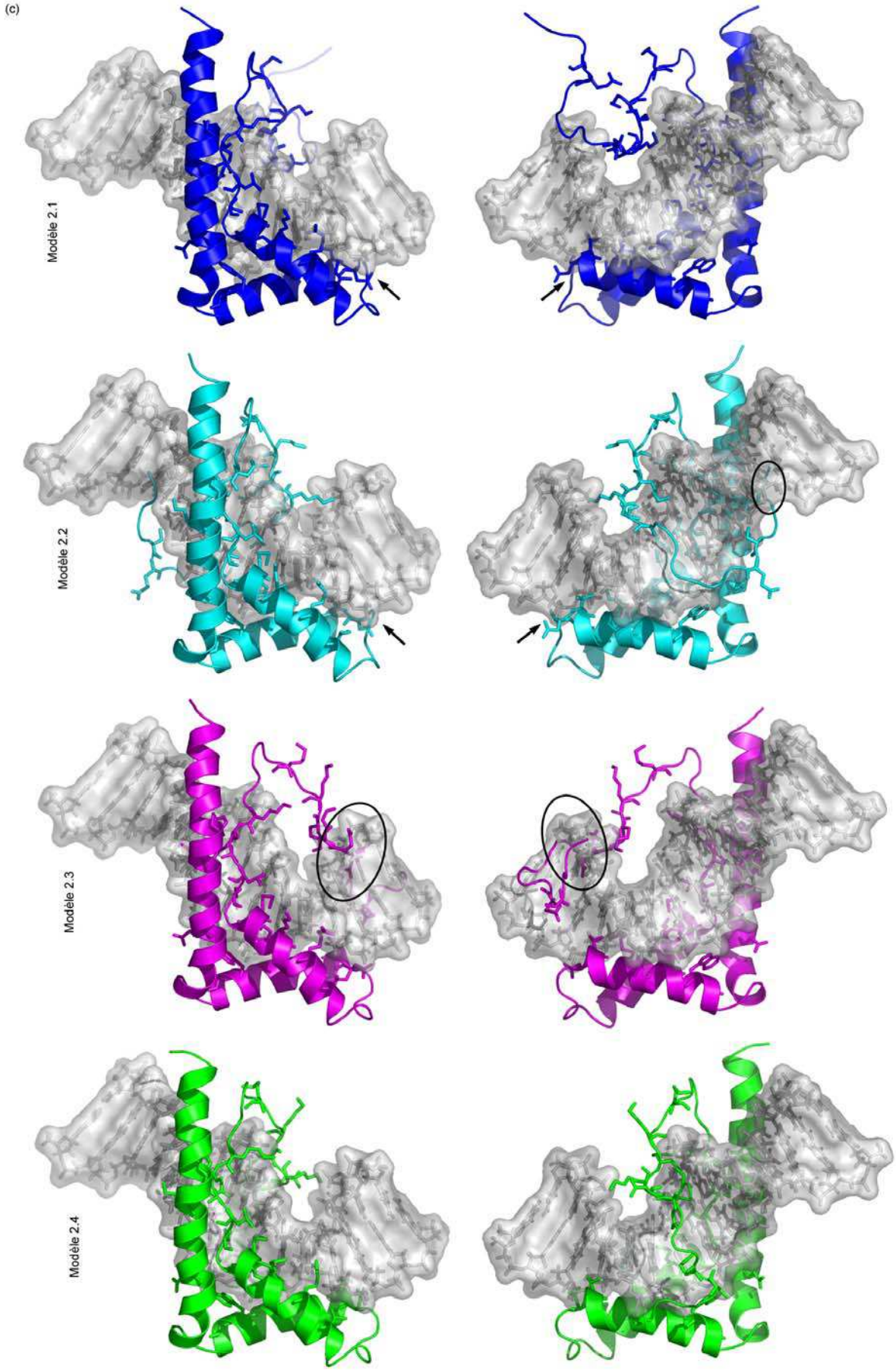


Figure 37. Structures support et modèles de PfHMGB1 & PfHMGB2 avec l'ADN.
Les chaînes latérales ne sont représentées que pour les acides aminés réellement liés à l'ADN pour NHP6A et par homologie avec la structure support pour les structures modèles. **(a)** Structure support NHP6A. **(b)** Structures modèles de PfHMGB1. Les résidus Val-8 et Lys-48 sont respectivement indiqués par des astérisques rouges et des flèches noires. **(c) Sur la page suivante.** Structures modèles de PfHMGB2. Les carbones α en conflit avec l'ADN sont entourés ou indiqués par des flèches.



(Suite de la Figure 37)

L'ADN sur lequel est fixé la protéine NHP6A présente une courbure de presque 90°. Le domaine 'HMG-box' de la protéine, comprenant les 3 hélices α, est fixé dans le petit sillon de l'ADN tandis que la partie N-terminale de la protéine, essentiellement basique et donc chargée positivement, se fixe dans le grand sillon comprimé sur la face de l'hélice opposée au petit sillon alors élargi (Figure 37a) [5, 96] de façon à stabiliser la courbure de l'ADN induite par le domaine 'HMG-box' [236, 298]. Toutes les structures modèles présentent un même agencement des hélices α dans le petit sillon de l'ADN et se différencient essentiellement par leur partie N-terminale.

Tous les modèles, excepté le modèle 2.4, ont un ou plusieurs carbones α en conflit avec la surface de l'ADN (indiqués par des flèches ou encerclés sur la Figure 37b et c) :

- la lysine de la boucle située entre les hélices α1 et α2 dans les structures modèles 1.1, 1.2, 2.1 et 2.2,
- les deux premiers acides aminés du modèle 2.2,
- les résidus Leu-10, Lys-12 et Gln-13 du modèle 2.3.

En ce qui concerne l'agencement des chaînes latérales dans chaque modèle, toutes les structures modèles présentent des conflits entre certaines chaînes latérales et la surface de l'ADN ; les chaînes latérales concernées appartiennent la plupart du temps à des acides aminés qui sont censés interagir avec l'ADN (par homologie avec la protéine NHP6A, Figure 38) donc qui sont assez proches de la double hélice pour former des liaisons hydrogène avec elle. Et les acides aminés concernés sont souvent des acides aminés qui ont une chaîne latérale assez longue (lysine et arginine) et donc difficile à positionner.

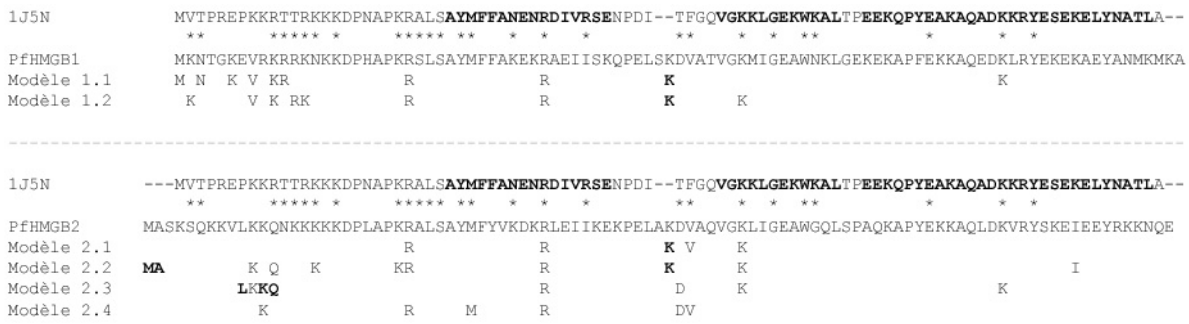


Figure 38. Résidus de chaque structure modèle en conflit avec la surface de l'ADN.
Pour la structure support 1J5N sont indiqués en gras : les résidus impliquées dans les 3 hélices α et par des astérisques (*) : les résidus en contact avec l'ADN. Pour les structures modèles, les résidus indiqués sont ceux qui entrent en conflit avec la surface de l'ADN soit par leur chaîne latérale, soit par leur carbone α (en gras).

Tout d'abord, en ce qui concerne PfHMGB1, certaines chaînes latérales sont en conflit dans les deux structures modèles : il s'agit des chaînes latérales des résidus Lys-10, Arg-23, Arg-36 qui devraient interagir avec l'ADN ainsi que celle du résidu Val-8 qui se situe dans la poche formée par la courbure de l'ADN indiqué par des astérisques rouges sur la Figure 37b. Les autres chaînes latérales en conflit avec la surface de l'ADN sont localisées dans la partie N-terminale de la protéine, au niveau du bras qui passe dans le grand sillon de l'ADN.

Pour ce qui est de PfHMGB2, les chaînes latérales des résidus Arg-26, Arg-39 et Lys-58 sont en conflit dans au moins 3 modèles : ces 3 résidus sont censés interagir avec l'ADN. Il semblerait que la trajectoire de la boucle située entre les hélices α1 et α2 soit meilleure dans les modèles 2.3 et 2.4 car seule la chaîne latérale du résidu Asp-52 entre en conflit avec la surface de l'ADN alors que dans les modèles 2.1 et 2.2, il s'agit du résidu Lys-51 tout entier. Les autres chaînes latérales présentant des conflits sont, comme dans le cas des modèles de PfHMGB1, situés dans la partie N-terminale de la protéine, au niveau du bras qui entoure la double hélice.

A la vue de ces résultats, une analyse plus poussée de la conformation des chaînes latérales s'est imposée et a été faite avec le programme SCit. Il en est ressorti qu'environ un tiers des chaînes latérales dans chaque structure modèle (respectivement 30, 28, 25, 22, 34 et 24 pour les modèles 1.1, 1.2, 2.1, 2.2, 2.3 et 2.4) se trouvaient dans une conformation inhabituelle pour la structure locale du squelette peptidique (données non présentées).

I.5 - Expériences biologiques

Toutes les analyses *in silico* amènent à penser que les facteurs PfHMGB1 & PfHMGB2 sont de vrais facteurs architecturaux de la famille HMGB et qu'ils sont alors capables de se lier à l'ADN et de le courber. Des expériences *in vitro* ont donc été pratiquées chez *P. falciparum* pour le démontrer.

En résumé, des analyses faites avec des protéines recombinantes ont permis de montrer les faits suivants : les deux protéines sont capables (i) d'interagir avec des structures d'ADN déformées (distordues) non conventionnelles, cette interaction étant plus efficace et spécifique pour PfHMGB1 que pour PfHMGB2, et (ii) de courber de l'ADN linéaire, encore une fois, PfHMGB1 étant plus efficace que PfHMGB2. Ces deux protéines ont été observées

dans les stades érythrocytaires asexués du parasite ainsi que dans les gamétocytes grâce à des expériences de Western-blot et d'immunofluorescence. Leur niveau d'expression est différent : en effet, la protéine PfHGMB1 est abondante dans les cellules au stade asexué alors que PfHMGB2 l'est dans les gamétocytes, ce qui est en corrélation avec les niveaux d'expression des transcrits, données obtenues grâce aux expériences de transcriptome à grande échelle effectuées dans le laboratoire d'Elizabeth Winzeler [224]. Non seulement le niveau d'expression des deux protéines est différent, mais leur localisation l'est aussi : tandis qu'on peut localiser les deux protéines dans le noyau des cellules asexuées et des gamétocytes, on trouve aussi PfHMGB2 dans le cytoplasme des gamétocytes.

Le fait qu'on puisse discriminer les deux protéines par leur affinité pour l'ADN, leur efficacité de liaison et de courbure de l'ADN, leur niveau d'expression (transcrit et protéine) ainsi que par leur localisation nous laisse penser que ces deux protéines n'ont pas un rôle redondant dans la cellule, notamment au niveau de la régulation transcriptionnelle des gènes impliqués dans le développement du parasite dans les érythrocytes.

Pour plus de détails sur les expériences biologiques menées sur ces deux facteurs architecturaux, vous pouvez vous reporter à l'article *High-Mobility-Group box nuclear factors of Plasmodium falciparum* (2006), situé à la fin de ce manuscrit

I.6 - Discussion & perspectives

Les protéines PfHMGB1 & PfHMGB2 de *Plasmodium falciparum* ne possèdent qu'un seul domaine 'HMG-box' comme les protéines de plantes et plusieurs protéines de levure et de drosophile. Les domaines 'HMG-box' des protéines parasitaires sont similaires aux domaines d'autres protéines eucaryotes (Figure 28 et Figure 30) et montrent la conservation de résidus caractéristiques importants pour la liaison à l'ADN et sa courbure, ce qui a déjà été montré pour la protéine HMG-D de drosophile par Frank Murphy IV et ses collaborateurs en 1999 [266]. La présence des résidus Ser-26 et Val-50 pour PfHMGB1 et Ser-29 et Val-53 pour PfHMGB2 (résidus encadrés sur la Figure 30 qui correspondent aux résidus Ser-10 et Val-32 de la séquence HMG-D) nous permet de classifier les protéines PfHMGB1 & PfHMGB2 dans la famille des protéines HMGB architecturales. Chez la drosophile, le résidu Ser-10 forme des liaisons hydrogène avec l'ADN alors que le résidu hydrophobe Val-32

s'intercale partiellement entre deux paires de bases, tout comme le résidu Met-13. Ces intercalations partielles des résidus Met-13 et Val-32, deux résidus situés au début des hélices α1 et α2 respectivement, introduisent deux vrilles successives dans l'ADN, ce qui accentue la courbure de l'ADN provoquée par l'élargissement du petit sillon lors de la liaison de la protéine à l'ADN, courbure qui peut atteindre un angle de 90° en un seul tour d'hélice. Comme les résidus Ser, Met et Val sont conservés chez PfHMGB1 & PfHMGB2, il semblerait logique de penser qu'ils ont la même fonction. En plus du domaine 'HMG-box', les protéines du parasite possèdent une petite région basique située en N-terminal du domaine 'HMG-box' mais n'ont pas de queue acide tout comme les protéines NHP6 de levure. Le domaine 'HMG-box' se lie à l'ADN dans le petit sillon et, dans le cas des protéines HMG-D et NHP6A, les régions basiques se lient sur la face opposée, c'est-à-dire dans le grand sillon [5, 96] de façon à stabiliser la courbure de l'ADN induite par le domaine 'HMG-box' [236, 298] et *in vitro* faciliter ainsi la circularisation de l'ADN [148, 412].

Les deux protéines plasmodiales ne présentant qu'un seul domaine 'HMG-box', nous avons voulu savoir si ce domaine était plus proche de la boîte A ou de la boîte B des protéines possédant deux domaines 'HMG-box' en tandem. Toutes les analyses mènent à la même conclusion : les domaines 'HMG-box' des deux protéines ressemblent plus à la boîte B. Tout d'abord, une analyse phylogénétique faite uniquement avec les domaines 'HMG-box' des protéines de *P. falciparum* et de divers vertébrés montre que les domaines des protéines du parasite se regroupent avec les boîtes B des vertébrés (voir l'article *High-Mobility-Group box nuclear factors of Plasmodium falciparum*, 2006). Ce résultat est conforté par l'observation suivante : alors que les boîtes de type B sont souvent trouvées seules dans les facteurs HMGB, les boîtes de type A sont généralement trouvées associées à une boîte de type B. De plus, il a été montré que la boîte B seule est capable de courber l'ADN *in vitro* et permet ainsi sa circularisation, alors que la boîte A, tout comme le domaine 'HMG-box' du facteur de transcription SRY, en est incapable [372]. Or il a été montré dans notre laboratoire que les protéines PfHMGB1 & PfHMGB2 sont capables de courber l'ADN (article *High-Mobility-Group box nuclear factors of Plasmodium falciparum*, 2006).

Deux modélisations par homologie ont été faites sur les facteurs PfHMGB. La première, une méthode « tout automatique », a modélisé les deux facteurs par homologie à la boîte B du facteur HMG1 de *Cricetulus griseus* (fichier PDB 1HSN, [317]) car c'est cette structure qui avait obtenu le meilleur score avec le programme TITO. Dans le domaine 'HMG-box', quatre hélices α ont été prédites (α1, α1', α2 et α3) se repliant en forme de L. Généralement, les domaines 'HMG-box' n'ont que trois hélices, mais l'échantillon utilisé pour obtenir la structure de la protéine par spectroscopie RMN contenait une molécule de βME liée au résidu Cys-14 situé dans la première hélice ; cela a eu pour effet de briser cette première hélice en hélices α1 et α1'. Même si quatre hélices α ont été prédites dans les deux protéines plasmodiales, leur position correspond aux hélices des autres domaines 'HMG-box' dont la structure est connue, comme Dm.HMG-D [266], Rn.HMG1_A [288] et Sc.NHP6A [5] (résidus soulignés dans la Figure 30). Néanmoins, ces structures modèles n'ont été construites qu'en remplaçant les carbones α de la structure support par les carbones α des séquences cibles : les deux résidus (KD) en plus dans les séquences cibles n'ont pas été modélisées ce qui explique que l'on obtienne un RMSD de 0,0 Å quand les structures modèles sont comparées à la structure support avec CE.

La deuxième modélisation a utilisé comme structure support la protéine NHP6A de *Saccharomyces cerevisiae* (fichier PDB 1J5N) [5]. Il semble que cette structure soit un meilleur support que celle utilisée précédemment, en partie à cause des résultats phylogénétiques obtenus avec la méthode de reconstruction UPGMA et du pourcentage d'identité assez élevé entre cette séquence support et les deux séquences cibles : NHP6A partage 45,36% et 45,45% d'identité respectivement avec PfHMGB1 et PfHMGB2, pourcentages qui augmentent à 52,11% et 50,70% quand on se restreint au domaine 'HMG-box' (Tableau 10). Ces pourcentages sont tout à fait corrects pour se lancer dans une modélisation par homologie. De plus, la structure de NHP6A a été obtenue alors que la protéine était liée à une double hélice d'ADN et cela apporte donc beaucoup de renseignements sur les résidus essentiels à la liaison à l'ADN. Sans oublier qu'en 1998, une équipe américaine a pu mettre en évidence, par délétion et mutagenèse, les résidus de NHP6A importants à son activité biologique (résidus encadrés sur la Figure 34) [412] :

- deux blocs de résidus basiques (KKR en positions 8-10 et RKKK en positions 13-16) sont nécessaires à la protéine, le deuxième pour l'efficacité d'interaction et de

courbure de l'ADN tandis que le premier permet de stabiliser la liaison ADN-protéine (ils font partie de la région basique déjà mentionnée sur la Figure 31) ;

- deux prolines sont importantes : la proline en position 21 contribue à la stabilité du repliement de la protéine, probablement par l'intermédiaire d'interactions hydrophobes avec des résidus situés à la fin de la troisième hélice α alors que la proline en position 18 semble faciliter le positionnement de la partie N-terminale de la protéine dans le grand sillon de l'ADN en infligeant un coude à la structure ;
- la tyrosine en position 28 et la phénylalanine en position 31 font partie de la première hélice α et leurs chaînes latérales sont orientées vers la deuxième hélice α avec laquelle elles interagissent pour stabiliser le repliement de la protéine et former un cœur hydrophobe ;
- la méthionine en position 29 semble jouer un rôle dans la courbure de l'ADN induite par NHP6A.

Tous les résidus importants pour l'activité de NHP6A sont présents dans les deux protéines plasmodiales (Figure 30 et Figure 34). Alors que la région basique est présente même si elle n'est pas identique, les autres résidus sont très bien conservés, la seule substitution ayant lieu dans la séquence de PfHMGB2 où le résidu Phe-31 est remplacé par une tyrosine, un acide aminé extrêmement proche, en terme de structure et de propriétés physico-chimiques, de la phénylalanine.

Deux et quatre modèles ont été sélectionnés respectivement pour PfHMGB1 et PfHMGB2, parmi les 100 modèles générés à chaque simulation. Tous ces modèles présentent trois hélices α se repliant en forme de L et se logeant dans le petit sillon de l'ADN. Ils diffèrent principalement par l'orientation prise par la partie N-terminale de la protéine et la boucle située entre la première et la deuxième hélices. Les trajectoires empruntées par la partie N-terminale des deux protéines sont toutes différentes les unes des autres mais aussi différentes de la trajectoire empruntée par la partie N-terminale de la structure support. Ceci peut s'expliquer par le fait que la partie N-terminale des protéines parasitaires est la partie qui diverge le plus de la séquence support. On peut noter entre autres l'absence de 2 prolines (Pro-4 et Pro-7 de NHP6A) dans les séquences plasmodiales, la proline étant un acide aminé particulier qui impose de fortes contraintes dans une structure du fait de sa flexibilité minimale.

Pour PfHMGB1, le programme Modeller a conservé la longueur de la première hélice et donc allongé la boucle alors que dans le modèle 1.2, la première hélice a été allongée et est « obligée » de se vriller pour que le repliement global de la protéine soit conservé (Figure 35c). C'est pourquoi ma préférence va au modèle 1.1 même si le carbone α du résidu Lys-48 est en contact avec la surface de l'ADN. De plus la partie N-terminale de la protéine semble, dans ce modèle, se positionner plus correctement dans le grand sillon de l'ADN (Figure 37b) même si ce positionnement n'est pas optimal. Quant à PfHMGB2, les 4 modèles sont très proches en ce qui concerne les hélices α. Néanmoins, les modèles 2.1 et 2.3 sont à éliminer de suite à cause de la partie N-terminale de la protéine (Figure 37c) : dans le premier cas, cette partie n'encercle pas du tout l'ADN, alors qu'elle est censée se loger dans le grand sillon de l'ADN pour stabiliser le complexe ADN-protéine et la courbure de l'ADN et dans le deuxième cas, cette partie de la protéine traverse la double hélice d'ADN. La boucle située entre la première et la deuxième hélices semble correctement modélisée dans le modèle 2.4 alors que le carbone α du résidu Lys-51 entre en conflit avec l'ADN dans le modèle 2.2 (Figure 36c). Cependant la partie N-terminale de la protéine parait mieux positionnée dans le grand sillon de l'ADN dans le modèle 2.2 que dans le modèle 2.4, si on ne tient pas compte des deux premiers acides aminés du modèle 2.2. Peut-être faudrait-il générer un plus grand nombre de modèles avec Modeller pour obtenir une structure modèle qui allierait le repliement de la partie N-terminale du modèle 2.2 (à l'exception de Met-1 et Ala-2) et le repliement de la fameuse boucle du modèle 2.4 ? Mais il faut aussi garder en mémoire qu'une hélice α a été prédite par le programme CONSENSUS dans la partie N-terminale de PfHMGB2 (Figure 34) et que celle-ci pourrait aussi se loger dans le grand sillon de l'ADN et interagir avec celui-ci pour stabiliser le complexe nucléoprotéique et la courbure de la double hélice. Une fois cette structure modèle obtenue, il sera indispensable de l'affiner de façon à ce qu'aucune chaîne latérale n'entre en conflit avec la surface de l'ADN. En effet, le programme Modeller n'est pas capable de prendre en compte l'ADN pour positionner les chaînes latérales des acides aminés, c'est pourquoi dans les six modèles étudiés, on trouve certaines chaînes latérales en conflit avec la surface de l'ADN.

L'analyse de la conformation des chaînes latérales par rapport à la conformation du squelette peptidique autour du résidu étudié avec le programme SCit a révélé que dans toutes les structures modèles, environ un tiers des chaînes latérales adoptent une

conformation inhabituelle. Néanmoins, l'analyse des chaînes latérales de NHP6A révèle que la moitié des chaînes latérales de la structure support se trouve aussi dans une conformation inhabituelle. Cet état est sûrement dû au fait que la protéine est fixée sur une double hélice d'ADN et que cela induit des contraintes sur toute la protéine.

Le remplacement des chaînes latérales en conformation inhabituelle par d'autres rotamères statistiquement plus probables n'a en rien arrangé les structures car, comme Modeller, SCit ne tient pas compte de la présence de l'ADN pour repositionner les chaînes latérales. De plus, le programme ne tient pas compte de la structure dans sa totalité, ce qui fait qu'après traitement par SCit, les structures ont des chaînes latérales en conflit avec d'autres chaînes latérales. Néanmoins, il est à noter que la structure support utilisée est une structure obtenue par spectroscopie RMN, et que même si la meilleure structure, selon les auteurs, a été utilisée ici, il en reste 19. Peut-être faudrait-il modéliser PfHMGB1 et PfHMGB2 à partir de plusieurs modèles provenant de la spectroscopie RMN et faire ensuite une dynamique moléculaire pour voir comment se comportent les protéines ? Cependant cette méthode est encore assez coûteuse en terme de temps de calcul et d'analyse pour un complexe ADN-protéine. Et il existe très peu de complexes ADN-protéine dans la PDB : au 4 avril 2006, 1 475 structures de la PDB sont des complexes nucléoprotéiques, toutes techniques expérimentales confondues, sur 35 917 structures répertoriées.

Même si les structures modèles sont encore à améliorer, les résultats de modélisation par homologie permettent d'avoir une idée assez précise de la structure de PfHMGB1 & PfHMGB2. Pour avoir encore plus de précisions, il faudrait répéter les expériences de mutagenèse faites par Yi-Meng Yen [412] pour voir si les effets remarqués avec NHP6A se retrouvent avec les protéines de *Plasmodium falciparum*.

Toutes les analyses *in silico* suggèrent que les protéines PfHMGB1 & PfHMGB2 sont de réels facteurs architecturaux capables de se lier à l'ADN et de le courber. Ceci a été validé par les analyses *in vitro* qui ont montré que les protéines recombinantes étaient capables d'interagir avec des structures d'ADN distordues et de courber l'ADN linéaire (voir l'article *High-Mobility-Group box nuclear factors of Plasmodium falciparum*, 2006). Il serait donc assez raisonnable de penser que les deux protéines plasmodiales jouent un rôle dans le remodelage

de la chromatine nécessaire aussi bien à la transcription qu'à la recombinaison ou la réplication de l'ADN. Chez les eucaryotes, un mécanisme d'action parmi d'autres a été proposé : les facteurs HMGB nucléaires modifieraient la structure des nucléosomes et ainsi relâcheraient l'ADN enroulé de façon à le rendre plus accessible aux complexes de remodelage et ainsi faciliter l'interaction entre les facteurs de transcription et leurs séquences cibles [380]. On a aussi observé une certaine relation entre les facteurs HMGB et l'histone H1 : elle est à l'origine de la balance entre les différentes conformations que peut adopter la chromatine, l'histone H1 compactant l'ADN à l'inverse des facteurs HMGB. Néanmoins, l'histone H1 n'a toujours pas été identifiée dans le génome de *Plasmodium falciparum* mais 60% des 5 300 gènes identifiés dans le génome du parasite n'ont toujours pas de fonction assignée. De plus, bien que les histones H1 soient très bien conservées chez les métazoaires, elles sont plus divergentes chez les protistes [295] et montrent donc une conservation plus faible que les histones H2A, H2B, H3 et H4 qui forment le cœur du nucléosome [15]. Certains protistes ont une protéine basique, riche en lysine, dont la composition ressemble à celle des histones H1 des eubactéries et au domaine C-terminal des histones H1 de plantes et d'animaux [196]. Peut-être existe-t-il dans le génome de *P. falciparum* un gène codant une protéine similaire aux histones H1 de protistes qui reste à ce jour non identifié ?

Pour finir, bien qu'observées toutes les deux dans tous les stades érythrocytaires du développement parasitaire, les protéines PfHMGB1 & PfHMGB2 présentent des niveaux d'expression différents, ce qui nous laisserait penser que PfHMGB1 pourrait être impliqué plutôt dans la prolifération du parasite tandis que PfHMGB2 le serait dans sa différenciation en gamétocytes. Ces deux protéines n'auraient donc pas une fonction redondante au sein du parasite, à l'image des protéines NHP6A et NHP6B chez *Saccharomyces cerevisiae* ; en effet, ces deux protéines sont très proches en longueur et identiques à 80% sur la totalité de leurs séquences mais ont des rôles différents.

De plus, les deux protéines n'interagissent pas avec l'ADN ni ne le courbent avec la même efficacité. Comme la topologie des trois hélices αest similaires dans les deux protéines, cette différence d'activité peut provenir de la partie N-terminale, partie la plus divergente entre les deux protéines.

II - Facteurs se fixant sur la boîte CCAAT

Les promoteurs eucaryotes sont caractérisés par deux types de motifs : les motifs spécifiques et les motifs « constitutifs » (voir p. 45 et 60). Les premiers modulent l'expression de gènes spécifiques, qui doivent s'exprimer à un moment clé du développement ou du cycle cellulaire ou encore en réponse à un stress ou à un stimulus. Les seconds sont impliqués dans l'activation des tous les promoteurs [296]. Un nombre limité de séquences, telles que les boîtes CCAAT ou GC, est présent dans quasiment tous les promoteurs et est reconnu par des facteurs bien connus aujourd'hui.

Il existe de nombreuses protéines se liant à l'ADN isolées et caractérisées qui comportent dans leur acronyme le mot CCAAT comme CTF/NF1 (CCAAT Transcription Factor/Nuclear Factor 1) ou CDP(CCAAT Displacement Protein).

Ces protéines reconnaissent des séquences palindromiques qui diffèrent du consensus de la boîte CCAAT établi par Philipp Bucher en 1990 [43] car ce consensus ne présente pas d'axe de symétrie. Seul le facteur appelé NF-Y, CBF ou encore HAP selon l'organisme, se fixe sur le pentanucléotide CCAAT et ne requiert aucun autre nucléotide [94, 327]. Et c'est sur ce facteur que je me suis focalisée.

Le facteur NF-Y appartient à la classe des protéines à motif à architecture β en contact avec le petit sillon de l'ADN. En réalité, il s'agit d'un facteur hétérotrimérique composé de trois sous-unités : NF-YA (CBF-B ou HAP2), NF-YB (CBF-A ou HAP3) et NF-YC (CBF-C ou HAP5), toutes essentielles pour la liaison à l'ADN [258, 344]. Un alignement des séquences protéiques provenant de différents espèces a montré que chaque sous-unité contient un domaine très conservé au cours de l'évolution [233]. Ces domaines conservés, comme pour tout facteur de transcription, sont indispensables pour la liaison à l'ADN mais sont aussi importants pour la trimérisation. Les sous-unités NF-YB et NF-YC forment un dimère très uni, auquel vient par la suite s'associer la sous-unité NF-YA. Le trimère résultant peut alors se lier à l'ADN avec une très grande spécificité et une très grande affinité [29, 202].

Dans le cas de la sous-unité NF-YA, le domaine caractéristique est répertorié dans la base de données Pfam sous le nom 'CBFB_NFYA' (Pfam : PF02045). Mais en ce qui concerne les sous-unités NF-YB et NF-YC, leurs domaines sont très proches en terme de séquences et

correspondent à un seul et unique domaine dans la base de données Pfam : 'CBFD_NFYB_HMF' (Pfam : PF00808).

Alors que le génome nucléaire n'était pas encore entièrement séquencé, les sous-unités du facteur NF-Y ont été recherchées chez *Plasmodium falciparum*. Cinq phases ouvertes de lecture ont été identifiées dans les contigs grâce au consensus du domaine caractéristique de chacune des trois sous-unités. Ces phases ouvertes de lecture ont été appelées NF-Y1 à NF-Y5 (Tableau 12). Chacune des séquences a été identifiée par un seul des trois consensus utilisés, il n'y a pas eu de résultats croisés : le consensus fait à partir des domaines caractéristiques des sous-unités NF-YB n'a pas donné les mêmes résultats que le consensus fait à partir des domaines caractéristiques des sous-unités NF-YC alors que ces domaines sont très proches.

Tableau 12. Cinq phases ouvertes de lecture ont été identifiées par homologie au consensus du domaine de liaison à l'ADN de chaque sous-unité du facteur NF-Y, avant le séquençage complet de *Plasmodium falciparum*.

Protéine	Chromosome	Taille	Consensus utilisé	Domaine caractéristique	Localisation cellulaire
NF-Y1	10	444 nt 147 aa	NF-YB	92 - 135	nucléaire (30%)†
NF-Y2	13	459 nt 152 aa	NF-YB	39 - 84	cytoplasmique (65%)
NF-Y3	11	1059 nt 352 aa	NF-YA	132 - 160	nucléaire (94%)
NF-Y4	14	3225 nt 1074 aa	NF-YC	33 - 100	cytoplasmique (65%)
NF-Y5	11	3561 nt 1186 aa	NF-YB	1132 - 1186	nucléaire (98%)

† Le programme PSORT a localisé la protéine NF-Y1 dans le noyau avec une certitude de 30% et dans la matrice mitochondriale avec une certitude de 10% quand l'origine de la séquence indiquée était « levure » ou « animal ». En revanche, le programme a localisé la protéine dans le stroma du chloroplaste avec une certitude de 88,2%, dans la membrane des thylacoïdes avec une certitude de 52,9%, dans les thylacoïdes avec une certitude de 52,9% et enfin dans le noyau de la cellule avec une certitude de 30%.

Pour NF-Y4, nous avons identifié une phase ouverte de lecture de 3 225 nucléotides sans intron codant une protéine de 1 074 acides aminés, alors que les programmes utilisés pour la prédiction de séquences codantes ont identifié une séquence codante de 3 231 nucléotides, composée de 2 exons (de 3 221 et 10 nucléotides) et codant une protéine de 1 076 acides aminés. Les deux annotations étaient donc très proches.

En revanche, pour NF-Y1 et NF-Y5, le résultat a été plus surprenant car il s'agissait en fait de la même protéine : NF-Y1 correspond à l'extrémité C-terminale de NF-Y5, alors que les deux phases ouvertes de lecture se situaient sur deux contigs censés appartenir à des chromosomes différents (Tableau 12) ! Cependant, il n'est pas rare que des erreurs se glissent dans les séquences accessibles au public avant la publication du génome, et ce, quel que soit le génome séquencé : c'est même pour cela que l'on parle de « séquences préliminaires ». Lors de la recherche de séquences codant des facteurs de transcription dans les contigs, il nous est arrivé d'identifier des protéines extrêmement bien conservées : il s'agissait en fait de protéines de levure car des morceaux de séquences de levure étaient restées parmi les séquences du parasite. Nous avons donc très tôt pris l'habitude de vérifier le pourcentage de A+T des séquences identifiées.

Pour le moment, aucune expérience biologique n'a été effectuée permettant de savoir quelles annotations sont correctes. L'annotation faite par le consortium a donc été conservée (Tableau 13), car elle repose sur la combinaison de plusieurs outils de prédiction de gènes.

Tableau 13. Quatre protéines pouvant intervenir dans la composition du facteur hétérotrimérique NF-Y ont été identifiées dans le génome nucléaire de *Plasmodium falciparum*.

Protéine	N° d'accession (PlasmoDB)	Taille	Domaine Pfam	Domaine caractéristique	Localisation cellulaire
NF-Y1	PF11_0477 chr. 11	3906 nt 1301 aa	CBFD_NFYB_HMF	1132 - 1186	nucléaire (98%)
NF-Y2	PF13_0043 chr. 13	459 nt 152 aa	CBFD_NFYB_HMF	39 - 84	cytoplasmique (65%)
NF-Y3	PF11_0204 chr. 11	1059 nt 352 aa	CBFB_NFYA	132 - 160	nucléaire (94%)
NF-Y4 (*)	PF14_0374 chr. 14	3231 nt 1076 aa	CBFD_NFYB_HMF	33 - 100	cytoplasmique (65%)

(*) Seule la protéine NF-Y4 possède un autre domaine putatif : un domaine IMP dehydrogenase / GMP reductase (IMP : Inosine-5'-MonoPhosphate, GMP : Guanosine 5'-MonoPhosphate)

Pour NF-Y3, d'après MotifScan, seule la moitié N-terminale du domaine caractéristique des sous-unités NF-YA serait présente. Cependant, lorsque la séquence NF-Y3 est alignée

avec les domaines 'CBFB_NFYA' de sous-unités NF-YA, dont certaines sont connues pour être fonctionnelles (Figure 39), même si le domaine 'CBFB_NFYA' n'est pas canonique chez NF-Y3, il semble bien présent avec quelques insertions. La présence d'insertions a d'ailleurs déjà été observée dans de nombreuses protéines chez *Plasmodium* [306], notamment dans les ARN polymérases I, II et III [123, 231, 232] et sera aussi observé dans PfMyb1 (voir p. 159).

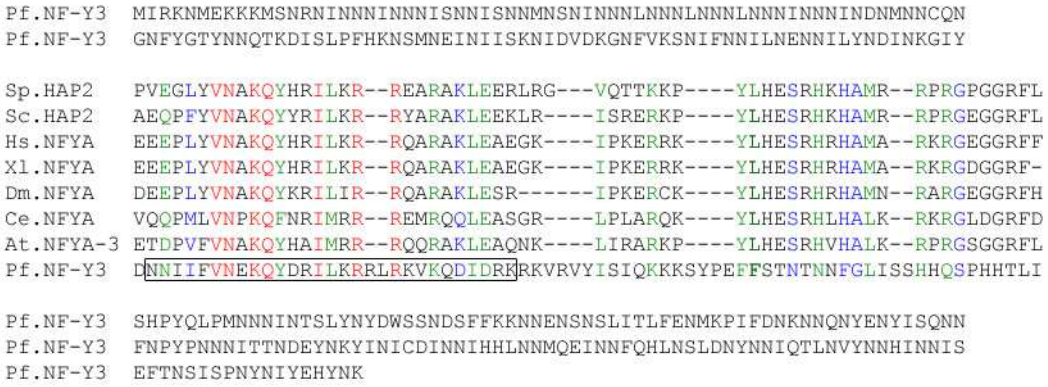


Figure 39. Alignement de la séquence complète de Pf.NF-Y3 avec les domaines 'CBFB_NFYA' de protéines NF-YA d'autres organismes eucaryotes.
Les couleurs correspondent à celles de ClustalW : en rouge, les identités ; en vert, les similitudes fortes ; en bleu, les similitudes faibles. Est encadré dans la séquence Pf.NF-Y3 le domaine 'CBFB_NFYA' identifié par MotifScan. Les noms des facteurs sont codés de la manière suivante : « abréviation_organisme.nom_facteur » (At, *Arabidopsis thaliana* ; Ce, *Caenorhabditis elegans* ; Dm, *Drosophila melanogaster* ; Hs, *Homo sapiens* ; Pf, *Plasmodium falciparum* ; Sc, *Saccharomyces cerevisiae* ; Sp, *Schizosaccharomyces pombe*).

Dans chacune des protéines NF-Y1, NF-Y2 et NF-Y3, le programme MotifScan a identifié un domaine 'CBFD_NFYB_HMF'. Même si ces séquences ont été obtenues par homologie à un seul des consensus utilisés au départ, une analyse phylogénétique a été faite pour savoir de quelle sous-unité se rapprochaient ces trois facteurs (Figure 40). Cette analyse a été faite avec les domaines caractéristiques de sept sous-unités NF-YB et sept sous-unités NF-YC (Annexe I.5).

D'après cette étude, même si les sous-unités NF-YB et NF-YC partagent le même domaine de la base de données Pfam, ils forment deux familles très distinctes. En ce qui concerne les facteurs plasmodiaux, il semblerait que le facteur NF-Y1 se rangerait du côté des sous-unités B tandis que le facteur NF-Y4 se rapprocherait des sous-unités C. Cependant, le facteur NF-Y2 ne se regroupe dans aucun des deux catégories, ce qui semble corroborer le fait que le programme MotifScan n'identifie pas dans NF-Y2 un domaine aussi long que dans les autres protéines.

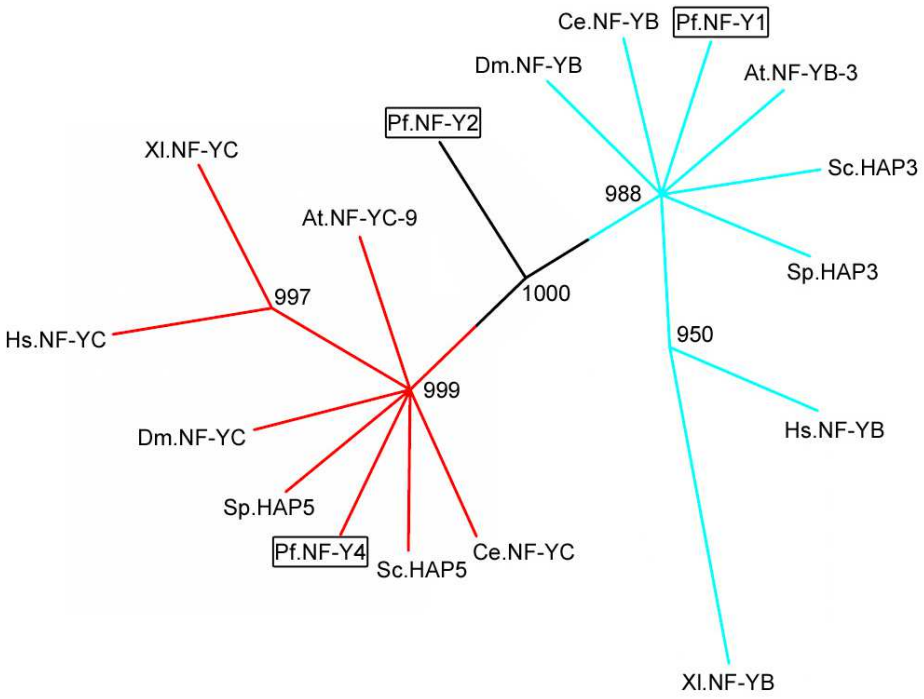


Figure 40. Arbre phylogénétique non enraciné construit avec la méthode Neighbor-joining à partir de 17 séquences représentant des sous-unités B et C du facteur NF-Y.
Les branches apparaissant dans moins de 600 arbres sur 1000 ont été fusionnées. A chaque nœud sont indiquées les valeurs de bootstrap (valeur maximale : 1000). En cyan, le groupe des sous-unités NF-YB ; en rouge, le groupe des sous-unités NF-YC. Les noms des facteurs sont codés de la manière suivante : « abréviation_organisme.nom_facteur » (At, *Arabidopsis thaliana* ; Ce, *Caenorhabditis elegans* ; Dm, *Drosophila melanogaster* ; Hs, *Homo sapiens* ; Pf, *Plasmodium falciparum* ; Sc, *Saccharomyces cerevisiae* ; Sp, *Schizosaccharomyces pombe*).

Il semblerait donc que le parasite *Plasmodium falciparum* renferme dans son génome des phases ouvertes de lecture codant toutes les sous-unités composant le facteur ubiquitaire NF-Y :

- une sous-unité A : Pf.NF-Y3 (PlasmoDB : PF11_0204),
- une sous-unité B : Pf.NF-Y1 (PlasmoDB : PF11_0477),
- une sous-unité C : Pf.NF-Y4 (PlasmoDB : PF14_0374).

Cependant, pour confirmer ces annotations, il sera indispensable d'étudier ces phases ouvertes de lecture de manière biologique : cloner les gènes, exprimer les protéines, regarder leur localisation cellulaire et surtout voir si les protéines forment non seulement un trimère, mais un trimère fonctionnel et efficace.

A l'heure actuelle, aucune expérience biologique n'a encore été conduite dans notre laboratoire pour valider les annotations de ces différents facteurs et, à ma connaissance, aucun autre laboratoire n'a publié d'articles concernant l'une ou l'autre de ces protéines. Néanmoins, il était important pour moi de savoir si de tels facteurs se fixant au promoteur proximal pouvaient exister chez *P. falciparum* car cela appuie encore plus l'idée que le parasite suit le mécanisme transcriptionnel général des eucaryotes.

III - Facteurs de transcription à doigts de zinc

Les doigts de zinc sont de petits domaines protéiques dans lequel le zinc a un rôle structural : il contribue à la stabilité du domaine. Ces domaines appartiennent à des protéines impliquées dans de nombreux processus cellulaires tels que la réplication et la réparation de l'ADN, la transcription et la traduction, le métabolisme et la signalisation, la prolifération cellulaire et l'apoptose. Il existe différentes classes de domaines en doigt de zinc (voir p. 66) parmi lesquelles la sous-classe C_2H_2. Rossella Tupler et ses collaborateurs [381] ont comparé les familles de facteurs de transcription chez l'homme, la drosophile, le nématode et la levure de boulanger : les facteurs contenant des doigts de zinc de type C_2H_2 (Pfam : PF00096) forment la famille la plus nombreuse (Figure 41). Lorsque nous avons commencé ce travail, aucun facteur de transcription n'était connu chez de *P. falciparum*. Avec le séquençage du génome, il s'est avéré que la famille des protéines en doigt de zinc était aussi une famille extrêmement représentée dans le parasite [68].

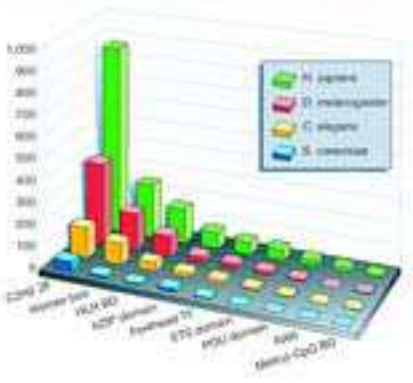


Figure 41. Comparaison des familles de facteurs de transcription chez les eucaryotes.
Les données proviennent d'une analyse des protéomes de l'homme, de la drosophile, du nématode et de la levure de boulanger selon la base de données INTERPRO qui comprend Pfam, PRINTS et PROSITE. Les familles de facteurs de transcription montrées ici sont les plus représentées. Image tirée de l'article de Tupler *et al.* [381].

Lorsque nous avons commencé à identifier des séquences codant des facteurs de transcription dans le génome de *P. falciparum*, le travail a été partagé. Parmi les facteurs possédant des domaines en doigt de zinc de type C_2H_2, ma directrice de thèse Catherine Vaquero a recherché des séquences pouvant coder des activateurs de type Krox. Quant à moi, de la même manière que j'avais cherché des facteurs se fixant à la boîte CCAAT, j'ai recherché le facteur Sp1, seul facteur connu pour se fixer à la boîte GC du promoteur proximal. Chacune de notre côté, nous avons trouvé une séquence contenant des doigts de zinc. Mais en fait, nous avions toutes les deux identifié la même séquence. Il fallait donc trancher.

Beaucoup de classifications ont été proposées pour les protéines possédant des domaines en doigt de zinc de type C_2H_2. En 1994, Pieler & Bellefroid ont divisé, de manière totalement subjective, cette classe de protéines en deux sous-classes [304]. D'un côté, les protéines telles que Egr1/Krox ou les protéines de la famille Sp, dont Sp1, qui ont dans leur séquence moins de cinq domaines C_2H_2. Ces protéines ont généralement été identifiées comme étant des activateurs ou des répresseurs transcriptionnels impliqués dans la régulation de la prolifération cellulaire et de la différenciation. De l'autre côté, les protéines qui ont cinq ou plus de cinq domaines C_2H_2. A part le facteur TFIIIA (co-facteur de l'ARN polymérase III) qui se fixe sur le gène codant l'ARNr 5S ainsi que sur l'ARNr 5S lui-même [373] et la protéine MZF1 qui régule la transcription du gène *cd34* [263], la ou les fonctions des protéines appartenant à cette classe restent encore inconnues à ce jour.

La protéine identifiée chez le parasite a une longueur de 1 461 acides aminés et contient neuf domaines C_2H_2 en C-terminal et aucun autre domaine caractéristique (Figure 42).

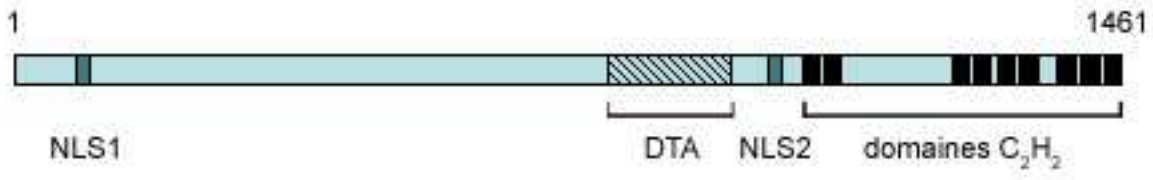


Figure 42. Schéma de la protéine PFL0465c.
La protéine de 1 461 aa contient neuf domaines en doigt de zinc de type C_2H_2 (en noir) tous situés dans la partie C-terminale de la protéine, une région de 160 acides aminés comportant douze séquences homologues au domaine de *trans*-activation (DTA) et deux signaux de localisation nucléaire putatifs (NLS1 et NLS2).

Elle est constituée d'un exon unique et l'annotation automatique faite lors du séquençage (PlasmoDB PFL0465c) correspond à la phase ouverte de lecture que nous avons annotée. D'après le programme PSORT, cette protéine serait une protéine nucléaire, ce qui confirmerait la présence des deux signaux de localisation nucléaire, qui ont été trouvés par MotifScan mais considérés comme statistiquement incertains.

Selon la classification de Pieler & Bellefroid, il semble donc peu probable que cette protéine soit une protéine Krox ou une protéine Sp1. Il est donc intéressant de savoir quelle protéine, ou quelle famille de protéines, contenant des domaines C_2H_2 se rapproche le plus de cette séquence. Pour cela, une base de données comprenant uniquement des protéines à domaines C_2H_2 (4 323 séquences) a été créée et une recherche de similitudes a été faite contre cette banque en utilisant comme séquence requête la partie C-terminale, c'est-à-dire la partie comprenant les neuf domaines C_2H_2 (421 aa). La séquence présentant le plus de similitudes avec la partie C-terminale, selon le programme BlastP, est la séquence du facteur TFIIIA du poisson-chat *Ictalurus punctatus* [286]. L'alignement obtenu avec BlastP, d'une longueur de 233 résidus, comporte 30% d'identités et 51% de similitudes ; il obtient un score de 125 et une E-value de 1e-29. La séquence de *P. falciparum* a donc été alignée avec les séquences de facteurs TFIIIA de divers organismes eucaryotes qui ont été étudiés moléculairement. Cet alignement a été fait en favorisant l'alignement des domaines C_2H_2 (Figure 43).

Le facteur TFIIIA du xénope est le premier facteur dans lequel des motifs en doigt de zinc ont été identifiés [159]. Ce facteur, comme ceux identifiés ultérieurement chez les autres eucaryotes, est uniquement requis pour la transcription par l'ARN polymérase III du gène codant l'ARNr 5S : il se fixe dans une région régulatrice interne du gène et recrute le facteur TFIIIC qui, lui, est indispensable pour le recrutement du facteur TFIIIB, complexe comportant trois protomères dont la TBP. Une fois le facteur TFIIIB intégré au complexe ADN-TFIIIA-TFIIIC, l'ARN polymérase III rejoint le complexe et commence la transcription du gène codant l'ARNr 5S [197]. Le facteur TFIIIA, en régulant la synthèse de l'ARNr 5S, régule donc indirectement la biogenèse des ribosomes.

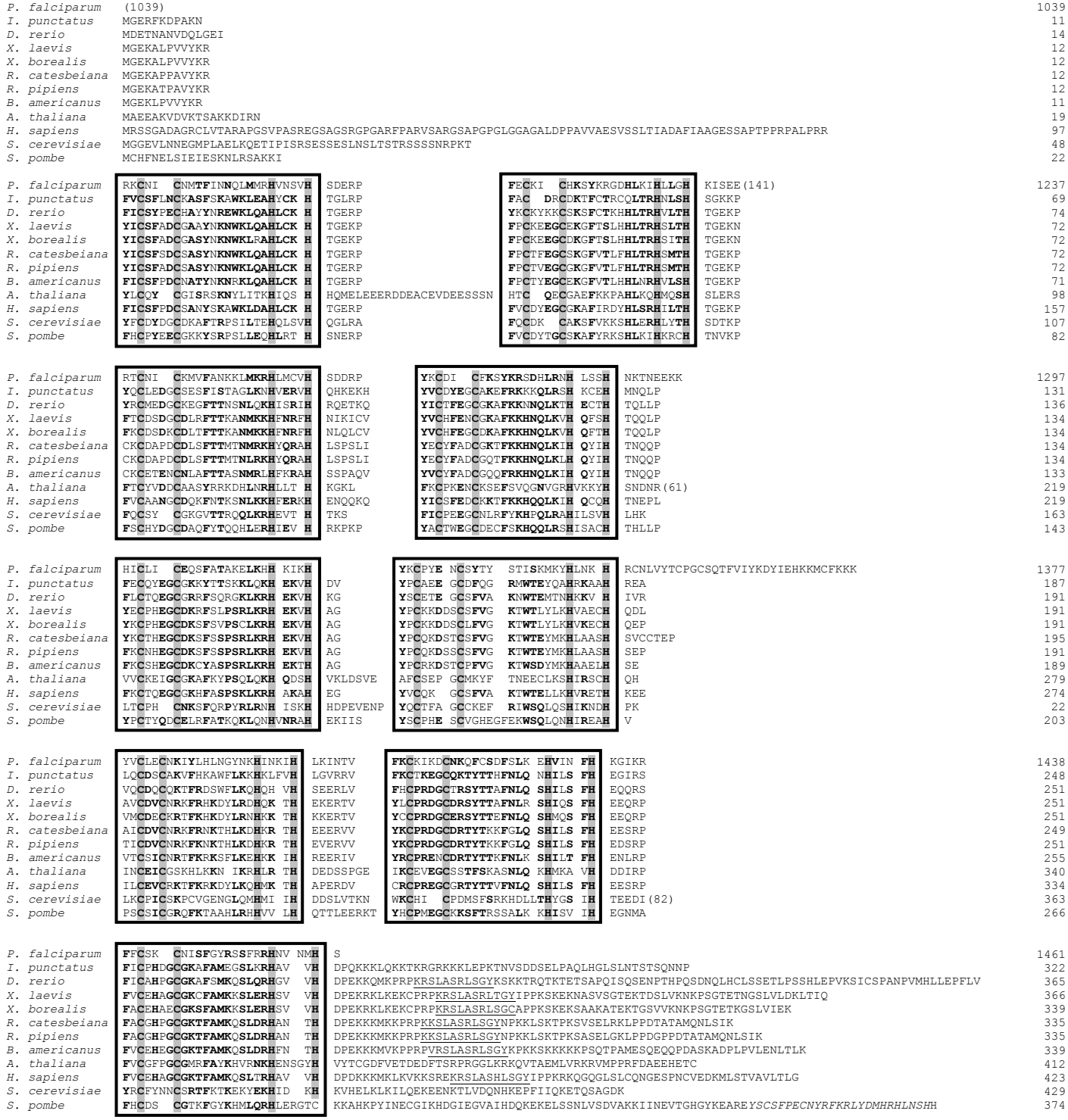


Figure 43. Alignement multiple de la partie C-terminale de la séquence identifiée chez *Plasmodium falciparum* avec les facteurs TFIIIA de divers organismes eucaryotes.

Dans les domaines C_2H_2 encadrés sont indiqués en gras les acides aminés identiques ou conservés dans plus de la moitié des séquences (résidus conservés : T/S, K/R, I/V, F/Y, L/M, N/D/E/Q). Les résidus Cys et His impliqués dans la liaison avec l'ion Zn^{2+} sont en plus surlignés en gris. Seule la protéine de *S. pombe* possède un dixième domaine C_2H_2 situé à l'extrémité C-terminale de la protéine et indiqué en italique. Les domaines de *trans*-activation sont soulignés. Les nombres entre parenthèses indiquent l'insertion de résidus qui ne sont pas montrés sur la figure. Les séquences utilisées ici ont été moléculairement étudiées et sont issues des poissons *Ictalurus punctatus* [286] et *Danio rerio* [8], des batraciens *Xenopus laevis* [140], *Xenopus borealis*, *Rana catesbeiana* [134], *Rana pipiens* et *Bufo americanus* [135], de la plante *Arabidopsis thaliana* [254], de l'homme [10], des levures *Saccharomyces cerevisiae* [12] et *Schizosaccharomyces pombe* [339]. La numérotation des séquences est indiquée à droite au bout de chaque ligne.

La séquence de *Plasmodium falciparum* semble donc avoir la même organisation que les facteurs TFIIIA car elle est constituée de 9 domaines C_2H_2 consécutifs. Elle présente tout de même une insertion de 141 acides aminés entre les domaines II et III ; mais la protéine de *A. thaliana* possède une insertion de 61 acides aminés entre les domaines IV et V et celle de *S. cerevisiae* une insertion de 82 acides aminés entre les domaines VIII et IX (chiffres entre parenthèses sur la Figure 43).

La seule grande différence qui existe entre la protéine de *P. falciparum* et les autres protéines se situe au niveau des parties N- et C-terminales. En effet, les facteurs TFIIIA ont une partie N-terminale « assez courte » allant d'une douzaine de résidus chez les poissons et les batraciens à une centaine de résidus chez l'homme et une queue C-terminale longue de 40 à 90 résidus. Chez *P. falciparum*, la protéine qui pourrait être un facteur TFIIIA possède une partie N-terminale longue de 1039 acides aminés et aucune queue C-terminale.

Les protéines des batraciens ainsi que celles du poisson-chat et de l'homme possèdent dans leur queue C-terminale une séquence d'une dizaine de résidus responsable de la *trans*-activation (résidus soulignés dans la Figure 43). Cette séquence a été recherchée dans la protéine du parasite grâce à un profil HMM généré à partir d'un alignement multiple des domaines de *trans*-activation. Le résultat a été étonnant : douze domaines, homologues au profil, ont été trouvés dans la séquence du parasite, tous regroupés dans une région d'environ 160 acides aminés (DTA sur la Figure 42). Le consensus du domaine de *trans*-activation identifié dans les séquences des batraciens, du poisson-chat et de l'homme et celui des domaines trouvés dans la séquence parasitaire sont assez proches (Figure 44).

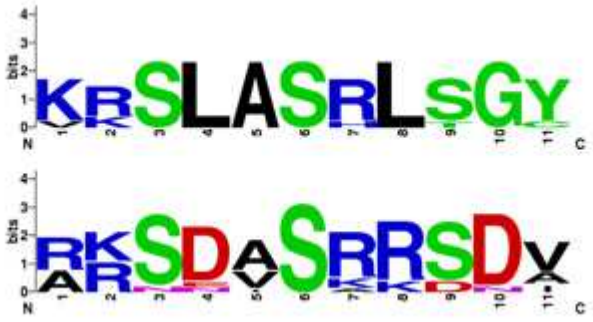


Figure 44. Séquences consensus du domaine de *trans*-activation présent dans différents facteurs TFIIIA (en haut) et d'un motif homologue répété dans la séquence de *P. falciparum* (en bas).
Le domaine de *trans*-activation est présent dans les facteurs TFIIIA de *D. rerio*, *X. laevis*, *X. borealis*, *R. catesbeiana*, *R. pipiens*, *B. americanus* et *H. sapiens*. Un domaine homologue au domaine de *trans*-activation est répété douze fois dans une région d'environ 160 acides aminés dans la séquence du parasite.

On pourrait les résumer de la manière suivante : des résidus basiques en positions 1 et 2, suivis d'une sérine et d'un résidu X ; un petit acide aminé comme l'alanine ou la valine en position 5 suivi d'une sérine puis d'un résidu basique et d'un résidu X ; enfin, une sérine en position 9 suivie de deux résidus X.

Il semblerait donc que *Plasmodium falciparum* possède donc une protéine ayant un nombre de domaines C_2H_2 et une organisation de ces domaines concordant avec ce que l'on trouve dans les facteurs TFIIIIA des divers eucaryotes. De plus, cette séquence possède des motifs qui pourraient jouer le rôle de *trans*-activateurs. Même si cette protéine est beaucoup plus longue que les facteurs TFIIIA eucaryotes, il serait intéressant de l'étudier car elle possède les deux domaines indispensables aux facteurs de transcription. Après avoir vérifié son annotation (un seul exon de 4386 pb), les propriétés biochimiques de la protéine pourront être examinées c'est-à-dire sa capacité à se fixer sur la région de contrôle interne du gène codant l'ARNr 5S, ainsi que sur l'ARNr 5S lui-même, et à promouvoir efficacement la transcription du gène codant l'ARNr 5S.

IV - Facteurs de transcription de la famille Myb

A la fin des années 1970, l'oncogène viral v-Myb est découvert chez le virus de la myéloblastose aviaire (AMV) [331] et la première séquence de sa contrepartie cellulaire chez le poulet, le proto-oncogène c-Myb, est publiée dès 1982 [145, 206]. A partir de ce moment-là, le nombre de protéines Myb annotées chez une grande variété d'eucaryotes, en allant du plus simple au plus complexe, va en grandissant ; tandis que leur étude fonctionnelle a mis en évidence que ces protéines étaient impliquées dans la régulation de la croissance cellulaire et de la différenciation, souvent en agissant de concert avec d'autres protéines se liant à l'ADN, la comparaison des séquences de ces protéines a permis d'identifier plusieurs domaines importants (Figure 45).

La comparaison des premières protéines homologues à c-Myb identifiées chez la souris, l'homme et le poulet [139, 146, 245, 329] montrent que les séquences sont très bien conservées dans leur ensemble. Dès 1983, Robert Ralston & J. Michael Bishop décrivent la région N-terminale de c-Myb [313] : celle-ci est constituée de 3 répétitions imparfaites (R1, R2 et R3),

chaque répétition étant constituée d'une cinquantaine d'acides aminés et suivant le schéma suivant : 3 tryptophanes conservés et régulièrement espacés de 18 ou 19 acides aminés [302]. Seule la protéine C1 du maïs ne répond pas tout à fait à cette description : en effet, celle-ci ne présente que deux répétitions, similaires à R2 et R3, et un des tryptophanes est remplacé par une isoleucine [299]. Néanmoins, en 1988, cette région est décrite comme étant une nouvelle structure de **domaine de liaison à l'ADN** ne ressemblant à aucune autre déjà connue [31]. De plus, des études ultérieures montrent d'une part que les répétitions R2 et R3 sont suffisantes pour que la protéine Myb se lie dans le grand sillon de l'ADN de manière séquence-spécifique [31, 242, 277] et d'autre part que R1 ne semble pas avoir d'interaction spécifique avec l'ADN [285] mais qu'elle augmenterait l'affinité de la protéine pour sa séquence cible [371] et la stabilité du complexe protéine-ADN [90]. La cristallographie aux rayons X et la spectroscopie RMN nous apprennent ensuite que chacune des répétitions se replie en 3 hélices α autour d'un cœur hydrophobe formé par les 3 tryptophanes, les 2 dernières hélices formant une structure en hélice-tour-hélice (Figure 17b) [125, 181, 284, 285].

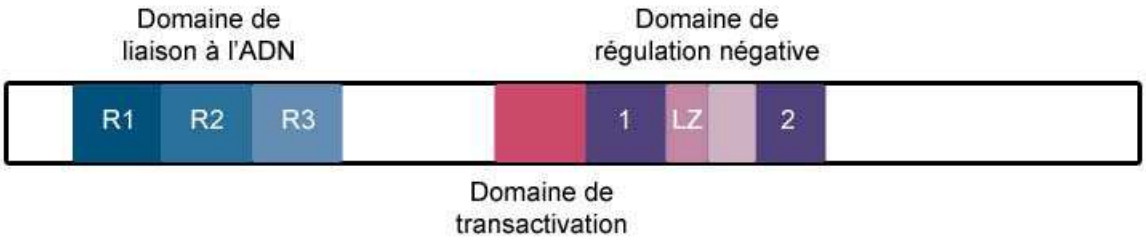


Figure 45. Représentation schématique des domaines fonctionnels de la protéine c-Myb de souris.
Le domaine de liaison à l'ADN est composé de trois répétitions imparfaites R1, R2 et R3 tandis que le domaine de régulation négative est composé de deux sous-domaines indépendants (1 et 2) entre lesquels se trouve une structure en forme de agrafe à leucines (LZ).

En plus du domaine de liaison à l'ADN, les protéines c-Myb présentent d'autres domaines fonctionnels :

- un domaine de *trans*-activation d'une cinquantaine de résidus, hydrophobe et légèrement acide, situé en C-terminal du domaine de liaison à l'ADN ;
- un domaine de régulation négative situé en C-terminal du domaine de *trans*-activation qui est constitué de deux sous-domaines capables d'inhiber l'activation transcriptionnelle indépendamment l'un de l'autre [97]. Entre ces deux sous-domaines non chevauchants se trouve une séquence formant une agrafe à leucines

(Figure 15a) qui permettrait la formation d'homodimères ou d'hétérodimères incapables de se lier à l'ADN [111, 191, 282].

Toutefois, les protéines identifiées ensuite chez la drosophile (D-Myb) [199, 302] et chez la levure (BAS1) [377], ainsi que les deux autres protéines identifiées chez l'homme (A-Myb et B-Myb) [281] montrent que c'est la partie N-terminale des protéines qui est la mieux conservée, à savoir le domaine de liaison à l'ADN, et que les deux derniers domaines décrits ne sont pas retrouvés dans toutes les protéines de type Myb. De plus, dans certains protéines Myb, comme trois des cinq protéines Myb de l'amibe *Dictyostelium discoideum* ou encore dans la protéine REB1 de *Kluyveromyces lactis* [264], ce domaine de liaison à l'ADN se trouve dans la partie C-terminale de la protéine, et non dans la partie N-terminale.

IV.1 - *P. falciparum* possède trois protéines avec plusieurs domaines Myb

Un consensus fait à partir de domaines de liaison à l'ADN de protéines Myb d'animaux, avec trois répétitions imparfaites, a permis d'identifier dans le génome de *P. falciparum* en cours de séquençage une phase ouverte de lecture de 1 245 nucléotides sur le chromosome 13 codant une protéine appelée PfMyb1 qui présentait des tryptophanes régulièrement espacés. Bien que la recherche ait été faite avec une séquence représentant trois domaines Myb, le programme MotifScan n'a reconnu dans la séquence plasmodiale qu'un seul domaine Myb (Pfam : PF00249) situé en C-terminal de la protéine.

P. falciparum présentant un génome particulièrement riche en A+T, cette protéine a été comparée aux domaines de liaison de protéines Myb appartenant à un eucaryote inférieur dont le génome est aussi très riche en A+T (77,2%) : l'amibe *Dictyostelium discoideum*. *D. discoideum* possède dans son génome cinq protéines Myb renfermant trois répétitions imparfaites en tandem (Tableau 14) : les trois premières ont été étudiées expérimentalement : DdMybA [358], DdMybB [292] et DdMybC [155], alors que les deux dernières, arbitrairement appelées DdMybD et DdMybE, ont été annotées lors du séquençage du génome [101].

Tableau 14. Récapitulatif des caractéristiques des protéines Myb de *D. discoideum*.

Protéine	N° d'accession (dictyBase)	Taille	Localisation des domaines Myb	Position du domaine de liaison à l'ADN
DdMybA	DDB0001463 chr. 6	1 230 aa	149 - 195 201 - 247 253 - 298	N-terminal
DdMybB	DDB0215356 chr. 2	711 aa	427 - 474 480 - 525 529 - 568	C-terminal
DdMybC	DDB0214816 chr. 3	580 aa	379 - 426 432 - 475 481 - 546	C-terminal
DdMybD	DDB0216340 chr. 6	577 aa	414 - 461 467 - 510 516 - 574	C-terminal
DdMybE	DDB0220517 chr. 1	669 aa	67 - 114 120 - 165 169 - 223	N-terminal

Seul le domaine Myb de PfMyb1 correspondant aux domaines R2 des séquences DdMyb avait été identifié par MotifScan. L'alignement entre la séquence entière de PfMyb1 et les domaines de liaison des cinq protéines de *Dictyostelium* (Figure 46) a donc permis d'identifier deux autres domaines situés de part et d'autre de celui qui avait déjà été identifié.

Les trois répétitions de PfMyb1 présentent toutefois quelques particularités. Le premier tryptophane du domaine R1 de PfMyb1 est remplacé par une tyrosine, tout comme le dernier tryptophane du domaine R2 ; quant au domaine R3, ce sont les premier et dernier tryptophanes qui sont remplacés par une phénylalanine. De plus, alors que les domaines R1 et R2 ont une taille en conformité avec les domaines Myb les plus couramment rencontrés, le domaine R3 présente une insertion de 6 acides aminés entre le tryptophane du milieu et la phénylalanine terminale, insertion que l'on retrouve aussi dans quatre des cinq protéines Myb de *D. discoideum*.

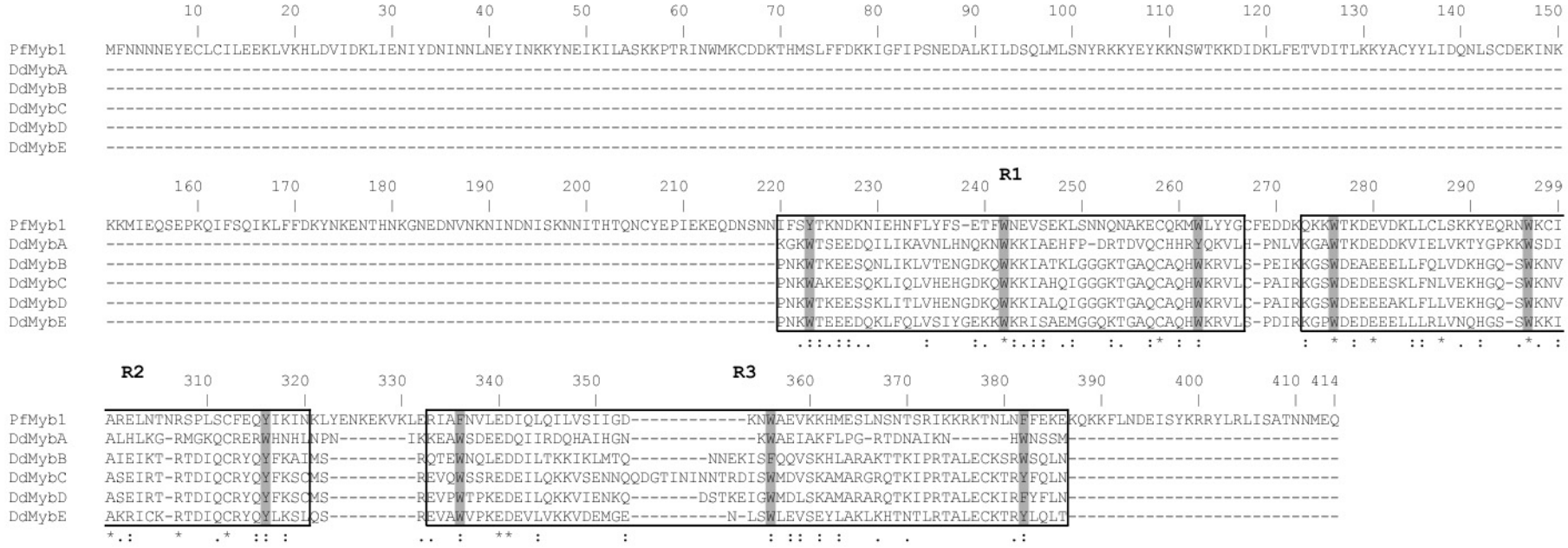


Figure 46. Alignement de la séquence complète de PfMyb1 avec les domaines de liaison à l'ADN des protéines Myb de *D. discoideum*.
Les trois répétitions imparfaites R1, R2 et R3 composant les domaines de liaison à l'ADN sont encadrées et numérotées. Elles ont toutes été identifiées par MotifScan en ce qui concerne les séquences de l'amibe *Dictyostelium discoideum* (DdMybA-E), mais seule R2 a été identifiée par ce programme dans le cas de PfMyb1. Les répétitions R1 et R3 de PfMyb1 ont été repérées par homologie aux répétitions R1 et R3 des séquences de *Dictyostelium*. Les résidus W, Y et F qui caractérisent les domaines Myb sont surlignés en gris. Les identités sont indiquées par une astérisque (*), les similitudes fortes par deux points (:) et les similitudes faibles par un point (.). La numérotation correspond à la séquence de PfMyb1.

En recherchant dans les bases de données de *P. falciparum* des séquences homologues au domaine R2 de PfMyb1, une autre phase ouverte de lecture a été identifiée : elle comporte 2 748 nucléotides, ne possède pas d'intron et code une protéine de 915 acides aminés possédant deux domaines Myb situés en N-terminal. Ces deux nouveaux domaines ont ensuite permis d'identifier une troisième phase ouverte de lecture de 7 737 nucléotides sans intron codant une protéine de 2 578 acides aminés. Cette dernière présente deux domaines Myb assez éloignés l'un de l'autre, l'un ayant été identifié par MotifScan et l'autre « à la main » (Tableau 15).

Tableau 15. Trois protéines contenant plusieurs domaines Myb ont été annotées dans le génome de *P. falciparum*.

Protéine	N° d'accession (PlasmoDB)	Taille	Localisation 'Domaine Myb'	Autres domaines	Localisation cellulaire
PfMyb1	PF13_0088 chr.13	1 245 nt 414 aa	220 - 266 273 - 320 333 - 386	Séquence de localisation nucléaire bipartite (putative)	nucléaire (94%)
PfMyb2	PF10_0327 chr.10	2 748 nt 915 aa	7 - 53 59 - 103	Séquence de localisation nucléaire bipartite (putative)	nucléaire (76%)
PfMyb3	PF10_0143 chr.10	7 737 nt 2 578 aa	656 - 702 1 588 - 1 639	Zinc finger domaine IMP dehydrogenase / GMP reductase	-

IMP : Inosine-5'-MonoPhosphate, GMP : Guanosine 5'-MonoPhosphate

Seule la protéine PfMyb3 possède, en plus des domaines Myb, un domaine en doigt de zinc de type ZZ (Pfam : PF00569) et un domaine 'IMP dehydrogenase / GMP reductase' (Pfam : PF00478). Alors que les protéines PfMyb1 et PfMyb2 auraient une séquence de localisation nucléaire qui n'est pas statistiquement sure, le programme PSORT leur prédit une localisation nucléaire. Malheureusement, comme pour PfHMGB3, le programme n'a pas fonctionné pour PfMyb3, à cause de la trop grande taille des protéines.

En octobre 2002, le séquençage du génome nucléaire du clone 3D7 étant terminé [132], nous avons pu vérifier que les phases ouvertes de lecture que nous avions déterminées étaient correctes.

Pour la suite de mon étude, je me suis intéressée uniquement à PfMyb1 car c'est la séquence qui a « le plus de chances » d'être un facteur de transcription de type Myb. En effet, elle présente trois domaines Myb et sa longueur semble être plus proche des premiers facteurs Myb identifiés.

IV.2 - Les domaines Myb de PfMyb1 présentent certains résidus conservés

L'étude et l'alignement des domaines Myb de protéines venant de divers eucaryotes au cours de l'évolution ont pu mettre en évidence en plus des trois tryptophanes initiaux qui peuvent à l'occasion être remplacés par les deux autres acides aminés aromatiques : la tyrosine et la phénylalanine, ou encore par l'isoleucine [235], la présence de nombreux résidus conservés dans toutes les répétitions (Figure 47) :

- un groupe d'acides aminés acides en C-terminal du premier tryptophane,
- une glycine en N-terminal du deuxième tryptophane,
- un groupe d'acides aminés aliphatiques en C-terminal du deuxième tryptophane,
- une arginine entre les deuxième et troisième tryptophanes.

Il existe un autre résidu hautement conservé dans ce domaine de liaison à l'ADN mais uniquement dans les répétitions R1 et R2 : une cystéine située entre les deuxième et troisième tryptophanes. Cet acide aminé, qui n'intervient pas dans la formation d'un pont disulfure, a une signification fonctionnelle importante, notamment celui de R2 : de son état oxydé ou réduit dépend un changement conformationnel de la protéine qui aura une influence sur la liaison de la protéine à l'ADN [268]. La substitution de la cystéine par une sérine dans c-Myb diminue significativement la liaison à l'ADN par déstabilisation de la structure de R2 [153, 268].

Les répétitions R1, R2 et R3 de PfMyb1 ont été alignées avec les répétitions d'autres protéines Myb, provenant d'eucaryotes supérieurs comme d'eucaryotes inférieurs, tout en considérant chaque répétition indépendamment des autres répétitions appartenant à la même protéine. Cela a permis de mettre en évidence que le domaine de liaison à l'ADN de PfMyb1 présentait bon nombre de résidus conservés (Figure 47).

Les domaines Myb des protéines d'eucaryotes inférieurs comme les levures *S. cerevisiae*, *K. lactis*, *S. pombe*, l'amibe *D. discoideum* et le parasite *P. falciparum*, ne « se comportent » pas tout à fait comme les domaines Myb des protéines d'eucaryotes supérieurs. En effet, les domaines Myb de chaque protéine ne se regroupent pas exactement avec les groupes R1, R2 et R3 des protéines Myb d'eucaryotes supérieurs.

```
MmC_R1        KTRWTREEDEKLKKLVEQNGT-----DDWKVIANYLPN-------RTDVQ-CQHRWQKVL
HsC_R1        KTRWTREEDEKLKKLVEQNGT-----DDWKVIANYLPN-------RTDVQ-CQHRWQKVL
XlC_R1        KTRWTREEDEKLKKLVEQNGT-----EEWKVIASFLPN-------RTDVQ-CQHRWQKVL
MmA_R1        RVKWTRDEDDKLKKLVEQHGT-----DDWTLIASHLQN-------RSDFQ-CQHRWQKVL
HsA_R1        RVKWTRDEDDKLKKLVEQHGT-----DDWTLIASHLQN-------RSDFQ-CQHRWQKVL
XlA_R1        KLRWTKDEDDKVKKLVEKHG------EDWGVVARHFIN-------RSEVQ-CQHRWHKVL
Dm_R1         GKRWSKSEDVLLKQLVETHG------ENWEIIGPHFKD-------RLEQQV-QQRWAKVL
MmB_R1        KVKWTHEEDEQLRALVRQFGQ-----QDWKFLASHFPN-------RTDQQ-CQYRWLRVL
HsB_R1        KVKWTHEEDEQLRALVRQFGQ-----QDWKFLASHFPN-------RTDQQ-CQYRWLRVL
XlB_R1        KVKWTPEEDETLKALVKKHGQ-----GEWKTIASNLNN-------RTEQQ-CQHRWLRVL
DdA_R1        KGKWTSEEDQILIKAVNLHNQ-----KNWKKIAEHFPD-------RTDVQ-CHHRYQKVL
ScBAS_R2      KGKWTQEEDEQLLKAYEEHG------PHWLSISMDIPG-------RTEDQ-CAKRYIEVL
DdC_R2        KGSWDEDEESKLFNLVEKHG------QSWKNVASEIRT-------RTDIQ-CRYQYFKSC
DdD_R2        KGSWDEEEEAKLFLLVEKHG------QSWKNVASEIRT-------RTDIQ-CRYQYFKSC
DdB_R2        KGSWDEAEEELLFQLVDKHG------QSWKNVAIEIKT-------RTDIQ-CRYQYFKAI
DdE_R2        KGPWDEDEEELLLRLVNQHG------SSWKKIAKRICK-------RTDIQ-CRYQYLKSL
Pf1_R2        QKKWTKDEVDKLLCLSKKYEQ-----RNWKCIARELNTN------RSPLS-CFEQYIKIN
SpCdc5_R2     KTEWSREEDEKLLHLAKLLP------TQWRTIAPIVG--------RTATQ-CLERYQKLL
DdC_R1        PNKWAKEESQKLIQLVHEHGD-----KQWKKIAHQIGGG------KTGAQ-CAQHWKRVL
DdD_R1        PNKWTKEESSKLITLVHENGD-----KQWKKIALQIGGG------KTGAQ-CAQHWKRVL
DdB_R1        PNKWTKEESQNLIKLVTENGD-----KQWKKIATKLGGG------KTGAQ-CAQHWKRVL
DdE_R1        PNKWTEEEDQKLFQLVSIYGE-----KKWKRISAEMGGQ------KTGAQ-CAQHWKRVL

MmC_R2        KGPWTKEEDQRVIELVQKYG----P-KRWSVIAKHLKG-------RIGKQ-CRERWHNHL
HsC_R2        KGPWTKEEDQRVIELVQKYG----P-KRWSVIAKHLKG-------RIGKQ-CRERWHNHL
MmA_R2        KGPWTKEEDQRVIELVQKYG----P-KRWSLIAKHLKG-------RIGKQ-CRERWHNHL
HsA_R2        KGPWTKEEDQRVIELVQKYG----P-KRWSLIAKHLKG-------RIGKQ-CRERWHNHL
XlC_R2        KGPWTKEEDQRVIELVHKYG----P-KRWSVIAKHLKG-------RIGKQ-CRERWHNHL
XlA_R2        KGPWTKEEDQRVIELVHKYG----P-KKWSIIAKHLKG-------RIGKQ-CRERWHNHL
MmB_R2        KGPWTKEEDQKVIELVKKYG----T-KQWTLIAKHLKG-------RLGKQ-CRERWHNHL
HsB_R2        KGPWTKEEDQKVIELVKKYG----T-KQWTLIAKHLKG-------RLGKQ-CRERWHNHL
XlB_R2        KGPWTKEEDEKVIELVKKYG----T-KHWTLIAKQLRG-------RMGKQ-CRERWHNHL
At1_R1        KGPWSKEEDDVLSELVKRLG----A-RNWSFIAR-SIP------GRSGKS-CRLRWCNQL
At2_R1        KGPWTEEEDAILVNFVSIHG----D-ARWNHIARSSGV------KRTGKS-CRLRWLNYL
ZmP_R1        RGRWTAEEDQLLANYIAEHG----E-GSWRSLPKNAGL------LRCGKS-CRLRWINYL
Zm1_R1        RGSWTPQEDMRLIAYIQKHG----H-TNWRALPKQAGL------LRCGKS-CRLRWINYL
Pp2_R1        RGPWTSEEDQKLVSHITNNG----L-SCWRAIPKLAGL------LRCGKS-CRLRWTNYL
KlREB_R2      RGKWTPEEDAELARWCAEK-----E-GQWSNIGKVLG--------RMPED-CRDRWRNYV
ScBAS_R1      RNSWSKDDDNMLRSLVNESA-(23)-IAWDVLATRFKHT-----VRTSKDV-RKRWTGSL
DdA_R2        KGAWTKDEDDKVIELVKTYG----P-KKWSDIALHLKG-------RMGKQ-CRERWHNHL
Dm_R2         KGPWTRDEDDMVIKLVRNFG----P-KKWTLIARYLNG-------RIGKQ-CRERWHNHL

MmC_R3        KTSWTEEEDRIIYQAHKRLG------NRWAEIAKLLPG-------RTDNA-IKNHWNSTM
HsC_R3        KTSWTEEEDRIIYQAHKRLG------NRWAEIAKLLPG-------RTDNA-IKNHWNSTM
MmA_R3        KSSWTEEEDRIIYEAHKRLG------NRWAEIAKLLPG-------RTDNS-IKNHWNSTM
HsA_R3        KSSWTEEEDRIIYEAHKRLG------NRWAEIAKLLPG-------RTDNS-IKNHWNSTM
XlA_R3        KSSWTEEEDRIIYSAHKRMG------NRWAEIAKLLPG-------RTDNS-IKNHWNSTM
XlC_R3        KSSWTEEEDRTIYEAHKRLG------NRWAEIAKLLPG-------RTDNA-IKNHWNSTM
MmB_R3        KSCWTEEEDRIICEAHKVLG------NRWAEIAKMLPG-------RTDNA-VKNHWNSTI
HsB_R3        KSCWTEEEDRIICEAHKVLG------NRWAEIAKMLPG-------RTDNA-VKNHWNSTI
XlB_R3        KSSWTEEEDRIICQAHKVLG------NRWAEIAKLLPG-------RTDNA-VKNHWNSTI
Dm_R3         KTAWTEKEDEIIYQAHLELG------NQWAKIAKRLPG-------RTDNA-IKNHWNSTM
At1_R2        RNSFTEVEDQAIIAAHAIHG------NKWAVIAKLLPG-------RTDNA-IKNHWNSAL
DdA_R3        KEAWSDEEDQIIRDQHAIHG------NKWAEIAKFLPG-------RTDNA-IKNHWNSSM
Pf1_R3        RIAFNVLEDIQLQILVSIIGD-----KNWAEVKKHMESLNSNTS-RIKKRKTNLNFFEKE
At2_R2        RGNITLEEQFMILKLHSLWG------NRWSKIAQYLPG-------RTDNE-IKNYWRTRV
Zm1_R2        RGNFTDEEEEAIIRLHGLLG------NKWSKIAACLPG-------RTDNE-IKNVWNTHL
ZmP_R2        RGNISKEEEDIIIKLHATLG------NRWSLIASHLPG-------RTDNE-IKNYWNSHL
Pp2_R2        RGIFSEAEENLILDLHATLG------NRWSRIAAQLPG-------RTDNE-IKNYWNTRL

KlREB_R3      ANKWSVEEEEKLKNVIHQMLDN-(54)-WTVVSEQMGGS------RSRIQ-CRYKWNKLL
DdC_R3        EVQWSSREDEILQKKVSENNQD----ISWMDVSKAMARGRQTKIPRTALE-CKTRYFQLN
DdD_R3        EVPWTPKEDEILQKKVIENKQDSTKEIGWMDLSKAMARARQTKIPRTALE-CKIRFYFLN
DdB_R3        QTEWNQLEDDILTKKIKLMTQNNEK-ISFQQVSKHLARAKTTKIPRTALE-CKSRWSQLN
ScBAS_R3      LREWTLEEDLNLISKVKAYG------TKWRKISSEMEF-------RPSLT-CRNRWRKII
DdE_R3        EVAWVPKEDEVLVKKVDEMGEN----LSWLEVSEYLAKLKHTNTLRTALE-CKTRYLQLT
SpCdc5_R1     GGAWKNTEDEILKAAVSKYGKN-----QWARISSLLVR-------KTPKQ-CKARWYEWI
Pf1_R1        IFSYTKNDKNIEHNFLYFSE------TFWNEVSEKLSNN------QNAKE-CQKMWLYYG
KlREB_R1      GKSFDESEEEALEQFIKE----------YQKIRG-LS--------RRQI--CERIWSNER
```

Figure 47. Alignement des répétitions R1, R2 et R3 de protéines Myb provenant d'organismes eucaryotes représentatifs (←).
Les répétitions ont toutes été considérées indépendamment des autres répétitions dans une même protéine. L'alignement a été fait avec ClustalW. Les domaines Myb sont regroupés en trois grandes familles selon les répétitions R1, R2 et R3 de la protéine c-Myb de souris. Les répétitions un peu plus distantes se distribuent entre ces trois familles. Les trois répétitions de PfMyb1 sont indiquées en gras. Chaque répétition est nommée de la façon suivante : l'abréviation du nom de l'organisme, suivie par une indication sur la protéine et par '_Rx' où x représente le numéro de la répétition. Les abréviations des noms des organismes sont indiquées en annexe (p. 231). Les tirets représentent les gaps et les nombres entre parenthèses indiquent l'insertion de résidus qui ne sont pas montrés sur la figure. Les résidus conservés sont surlignés : en gris, les tryptophanes (ou tyrosines, phénylalanines et isoleucines) caractérisant les domaines Myb ; en jaune, les résidus acides ; en magenta, la glycine ; en cyan, les résidus aliphatiques ; en vert, l'arginine ; en rouge, la cystéine des répétitions R1 et R2.

En ce qui concerne les trois domaines Myb de PfMyb1, tous les résidus conservés dans la plupart des domaines Myb ne sont pas retrouvés. Le domaine R1 ne possède pas d'arginine entre les deuxième et troisième tryptophanes ; les domaines R1 et R2 ne possèdent pas de résidu glycine entre les premier et deuxième tryptophanes. Mais à part ça, tous les autres résidus sont conservés : toutes les répétitions ont un ou plusieurs acides aminés acides après le premier tryptophane, un ou plusieurs acides aminés aliphatiques après le deuxième tryptophane. De plus, les répétitions R1 et R2 ont une cystéine juste avant le troisième tryptophane.

IV.3 - Chaque domaine Myb comporte plusieurs hélices

La structure du facteur PfMyb1 n'étant pas encore connue, nous avons donc décidé de modéliser la structure de ce facteur, ou tout du moins son domaine de liaison à l'ADN, par homologie à des structures connues (voir Méthodes p. 89).

La méthode « tout automatique » utilisée pour tenter de modéliser le facteur PfMyb1 a posé problème. L'analyse des structures des protéines Myb présentes dans la PDB a montré qu'aucune protéine Myb n'avait été cristallisée dans sa totalité mais que seule la structure du domaine de liaison à l'ADN (soit les trois répétitions, soit uniquement R2 et R3) était connue. La requête a donc été lancée sur le méta-serveur @TOME avec la séquence du domaine de liaison à l'ADN de PfMyb1 correspondant aux trois répétitions imparfaites identifiées (Figure 46). Néanmoins les résultats de la requête étaient déroutants. En effet, la structure

support ayant le score TITO le plus faible (fichier PDB : 1SED correspondant à une protéine putative de *Bacillus subtilis*), donc celle qui est considérée comme étant la meilleure pour modéliser le domaine de liaison à l'ADN de PfMyb1, n'est apparue dans les résultats que d'un seul des six programmes utilisés avec une E-value associée de 4,795 et l'alignement entre la séquence cible et la séquence support ne présentait que 11,6% d'identité. Ceci n'est donc pas très prometteur pour se lancer dans une modélisation par homologie car il est préférable d'avoir une E-value la plus proche de zéro possible et un pourcentage d'identité dans l'alignement supérieur à 25%. Comme ces résultats n'étaient pas très encourageants, je suis passée à la deuxième méthode de modélisation par homologie, à savoir en utilisant le programme Modeller. Il y a toutefois une chose à retenir des résultats obtenus : sur les vingt-deux structures sélectionnées par les programmes du méta-serveur @TOME et quel que soit le score TITO qui leur a été assigné, quatorze ont été classées dans la base de données CATH et toutes appartiennent à la classe « Mainly Alpha » et à l'architecture « Alpha Orthogonal Bundle ». Il y a donc de fortes chances pour que la protéine PfMyb1, ou tout du moins son domaine de liaison à l'ADN, comporte des hélices α formant un « paquet » dans lequel les hélices seraient perpendiculaires.

Les séquences les plus proches du domaine de liaison à l'ADN du facteur PfMyb1 (soit une séquence de 166 acides aminés) ont été recherchées dans la PDB. Les résultats obtenus sont regroupés dans le Tableau 16.

La première remarque à faire est que la nature des réponses n'est pas très variée. En effet, sur les sept structures dont la séquence est homologue au domaine de liaison à l'ADN de PfMyb1, cinq correspondent à la protéine c-Myb de souris, une à la protéine v-Myb du virus de la myéloblastose aviaire et enfin la dernière à la protéine B-Myb du poulet. Ensuite, le pourcentage d'identité entre la séquence requête et n'importe quelle séquence homologue est très faible, à la limite de la modélisation par homologie, cette limite étant située vers 25%. De plus, les protéines Myb sélectionnées sont réduites aux répétitions R2-R3, sauf dans le cas du fichier PDB 1H88 qui contient les trois répétitions du domaine de liaison à l'ADN de la protéine c-Myb murine. Mon choix se porte donc sur cette structure car elle fait partie des meilleurs pourcentages d'identité et que c'est la seule à avoir un domaine de liaison à l'ADN complet. Une dernière remarque est à faire : la structure 1SED qui avait été considérée

comme étant la meilleure structure support par le programme TITO n'apparaît pas dans les résultats de la recherche de séquences homologues au domaine de liaison à l'ADN de PfMyb1, et ce que l'on utilise BLASTP ou Psi-BLAST, alors que la structure 1H88 apparaissait dans les résultats du méta-serveur @TOME.

Tableau 16. Structures PDB dont la séquence est homologue au domaine de liaison à l'ADN du facteur PfMyb1.

N° accession PDB / SwissProt	Description	Méthode	Remarques	Données sur les résultats de la requête BlastP			
				Longueur	Score	% identité	E-value
1H89_C / P06876 (†)	*M. musculus* c-Myb	RX (2,45Å)	Gln77-Arg191 (R2-R3)	159	90	30	0,005
1H88_C / P06876 (†)	***M. musculus* c-Myb**	**RX (2,8Å)**	**Gly39-Arg190 (R1-R3)**	**159**	**90**	**30**	**0,005**
1A5J / Q03237	*G. gallus* B-Myb	RMN, 32	Gly1-Thr110 (R2-R3)	110	84	26	0,02
1MSF_C / P06876 (‡)	*M. musculus* c-Myb	RMN, 25	Met89-Val193 (R2-R3)	105	76	24	0,15
1MSE_C / P06876 (‡)	*M. musculus* c-Myb	RMN, 1	Met89-Val193 (R2-R3)	105	76	24	0,15
1GV2_A / P06876	*M. musculus* c-Myb	RX (1,68Å)	Glu89-Arg190 (R2-R3)	105	76	24	0,15
1H8A_C / P01104 (†)	*Virus AMV* v-Myb	RX (2,23Å)	Asn87-Arg191 (R2-R3)	128	74	23	0,25

La colonne 'Méthode' contient différentes informations : la technique utilisée pour obtenir la structure de la protéine (RMN ou RX pour cristallographie aux rayons X), suivie, dans le cas de la RMN, du nombre de structures dans le fichier PDB, ou dans le cas de la cristallographie, de la résolution du cristal. Dans la colonne 'Remarques' est indiqué la séquence exacte de la structure du fichier et ce à quoi elle correspond. La colonne 'Données sur les résultats de la requête BlastP' est séparée en quatre sous-colonnes : la longueur de l'alignement entre la séquence requête et la séquence homologue, le score, le pourcentage d'identité et la E-value associés à cet alignement.
(†) Complexes faits d'un double brin d'ADN de 26 pb sur lequel sont fixés la protéine C/EBPβ humaine et le domaine de liaison à l'ADN de c-Myb (R2-R3 pour 1H89 et R1-R3 pour 1H88) ou de v-Myb (R2-R3 pour 1H8A). (‡) Domaine de liaison à l'ADN de c-Myb (R2-R3) fixé à un double brin d'ADN de 16 pb, 1MSE étant jugée comme étant la meilleure structure de 1MSF.

Le pourcentage d'identité entre la séquence cible et la séquence support étant très faible, il a fallu préparer l'alignement nécessaire à Modeller avec une très grande minutie. Pour cela, plusieurs alignements entre la séquence cible et la séquence support (2 à 2, multiple et structural) ont été faits puis comparés pour en faire un « alignement consensus » auquel différentes informations comme les résidus en contact avec l'ADN et la position des structures secondaires vraies et prédites ont été intégrées. Il en est ressorti que le domaine de liaison à l'ADN de PfMyb1 pouvait être modélisé de deux manières différentes (Figure 48), même si la fin de la troisième répétition n'a, à chaque fois, pas pu être prise en compte à cause de l'insertion de six acides aminés dans la deuxième moitié de cette répétition. La première méthode consiste à modéliser le domaine en une seule fois grâce à un seul alignement ; la deuxième consiste à modéliser le domaine en deux fois grâce à deux

alignements chevauchants et à réunir, si possible, les deux sous-domaines grâce à leur partie commune. Cette deuxième méthode a pu être envisagée car les structures des différentes répétitions sont semblables et dans ce cas de figure, « découper l'alignement » augmente le pourcentage d'identité ce qui n'est pas négligeable ici. En effet, le pourcentage d'identité entre la séquence cible et la séquence support est légèrement en dessous de la limite acceptable pour faire de la modélisation par homologie (25%) mais lorsque les fortes similitudes sont prises en compte en plus des identités, les pourcentages dépassent 45%.

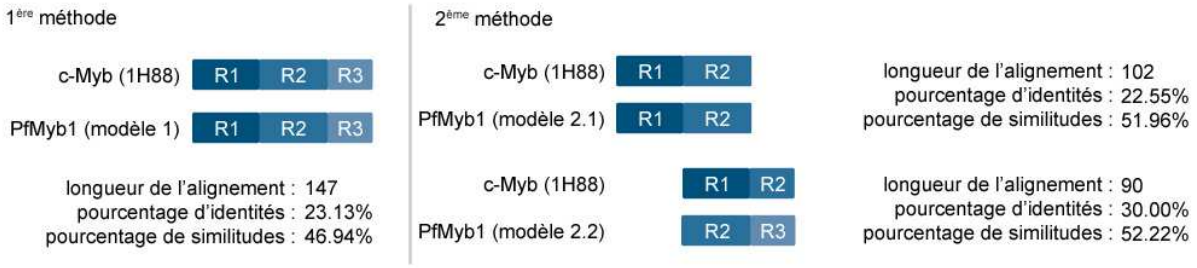


Figure 48. Schématisation des alignements utilisés pour la modélisation par homologie.
Le pourcentage de similitudes comprend les identités et les similitudes considérées comme fortes par ClustalW.

La position des structures secondaires prédites dans la structure support (Figure 49) est assez correcte mais leur longueur ne correspond pas toujours à la réalité. Les gaps ont tous été introduits en dehors des hélices vraies de la structure support sauf dans la première hélice de la répétition R1 où l'introduction d'un gap dans la séquence cible était nécessaire pour que tous les « tryptophanes » soient correctement alignés.

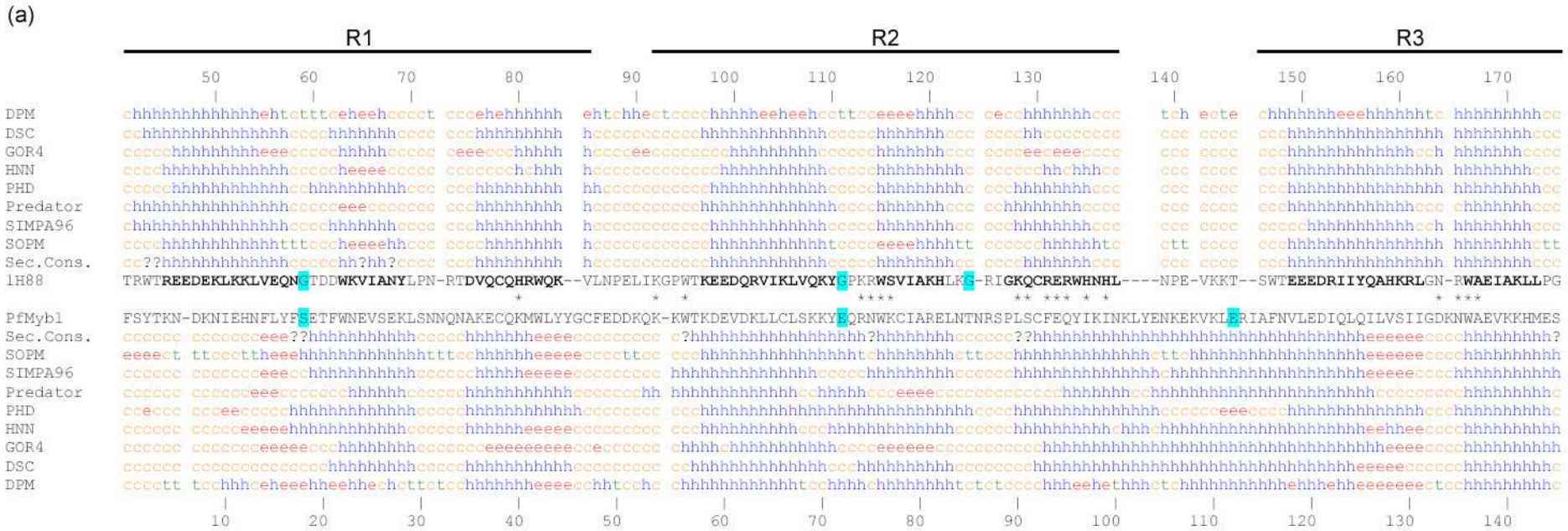
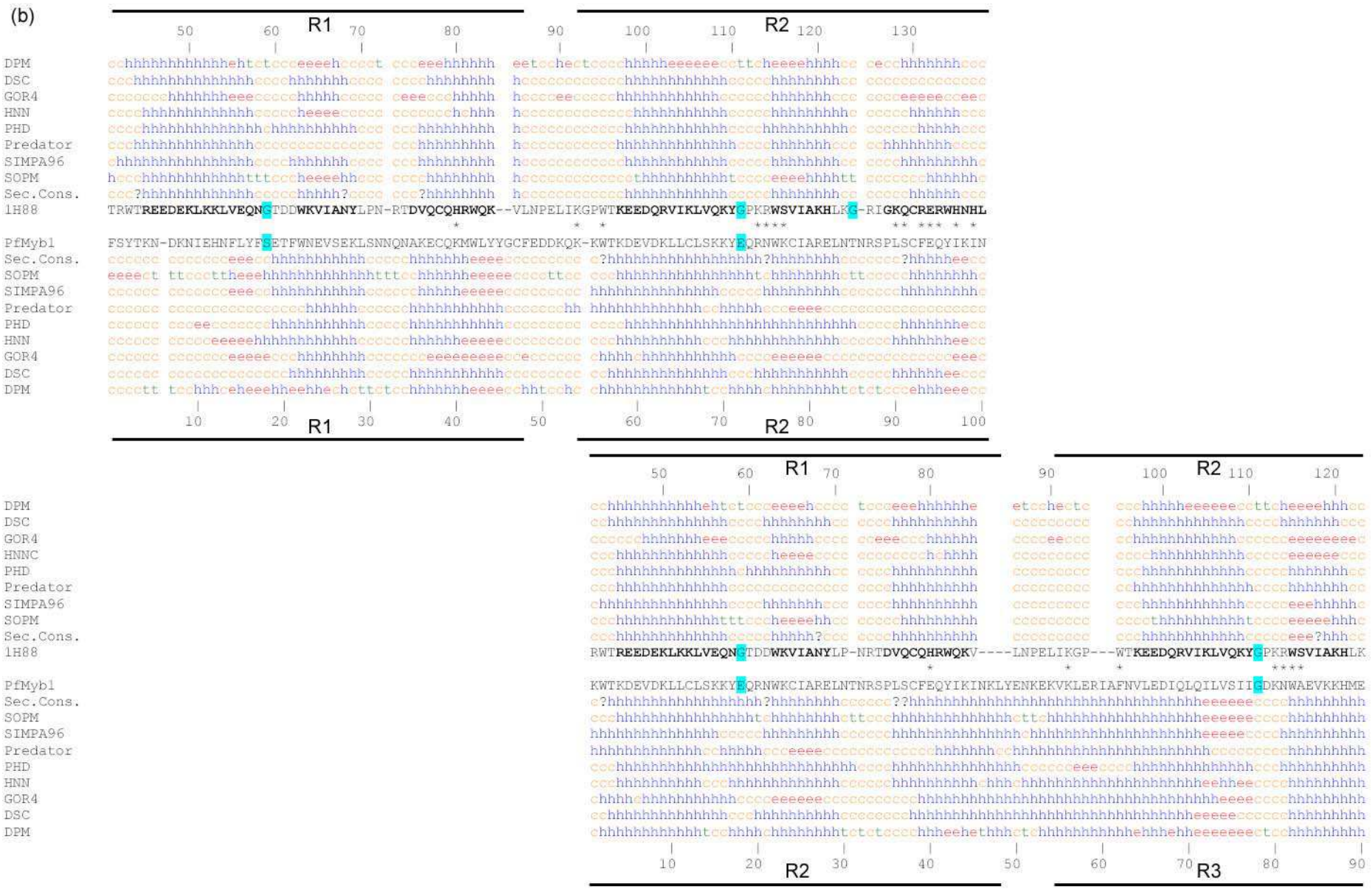


Figure 49. Alignements du domaine de liaison à l'ADN de PfMyb1 avec la séquence support c-Myb de souris (1H88).
(a) Alignement utilisé pour la première méthode de modélisation. **(b) Page suivante.** Alignements utilisés pour la deuxième méthode de modélisation.
En noir se trouvent la séquence cible PfMyb1 et la séquence de la structure support 1H88. La numérotation des séquences est faite en accord avec le fichier PDB pour la structure support et avec les fichiers générés par Modeller pour les structures modèles. Les caractères gras indiquent les hélices α réelles de la structure support et les astérisques (*) les résidus de c-Myb en contact avec l'ADN, comme indiqué dans le fichier PDB 1H88. En cyan sont surlignés les résidus situés en dehors des zones autorisées du diagramme de Ramachandran. Chaque séquence a été soumise au programme CONSENSUS qui regroupe différents outils de prédiction de structures secondaires. Sont représentés : en bleu les hélices, en rouge les brins, en vert les β-turns, en jaune les coils et par un ? les états ambigus dans le consensus des prédictions de structures secondaires.


(b)
R1
R2
R3
DPM
DSC
GOR4
HNN
HNNC
PHD
Predator
SIMPA96
SOPM
Sec.Cons.
1H88
PfMyb1

Parmi les 100 modèles générés par Modeller pour chacun des alignements, seul un modèle a été retenu à chaque fois grâce à la valeur de la fonction objective (voir p. 240).

Commençons par le premier modèle, appelé 'modèle 1', qui a été obtenu grâce à un seul alignement comprenant les répétitions R1 et R2 complètes et les deux premiers tiers de la répétition R3. Quand la structure modèle et la structure support sont alignées avec le programme CE, le RMSD est de 3,3 Å ce qui est assez élevé mais le Z-score est plutôt bon (5,7). La qualité du modèle 1 a été vérifiée (voir Figure 68, p. 241) : que ce soit avec Verify3D ou ProSa2003, il semble que les extrémités du modèle ne soient pas très correctes d'un point de vue conformationnel, ce qui peut être expliqué par le fait que nous en sommes en présence d'un domaine protéique et non d'une protéine entière. Les extrémités sont donc moins soumises aux contraintes que l'on pourrait rencontrer dans une protéine entière. Le diagramme de Ramachandran de la structure modèle indique que trois résidus sont situés en dehors des zones autorisées : dans l'alignement utilisé pour la modélisation, deux des résidus correspondants dans la structure support sortent aussi de ces zones (résidus surlignés en cyan sur la Figure 49).

Quand le modèle est superposé à la structure support (Figure 50a), on remarque tout de suite qu'il y a une petite hélice supplémentaire située juste après la troisième hélice de la répétition R2 (indiquée par une flèche). De plus, les répétitions ne peuvent pas être superposées toutes les trois en même temps : si les trois hélices de la répétition R2 sont très bien superposées, les hélices des deux autres répétitions ne le sont pas. Cependant, lorsque la structure modèle et la structure support sont découpées artificiellement pour isoler les trois répétitions et que ces trois répétitions sont superposées indépendamment les unes des autres (Figure 50b), les hélices se superposent très bien dans chacune des répétitions. Ce sont donc les boucles qui réunissent ces trois répétitions qui ont, dans la structure modèle, une trajectoire différente de ce que l'on peut observer dans la structure support. De ce fait, la position dans l'espace des trois répétitions les unes par rapport aux autres est différente.

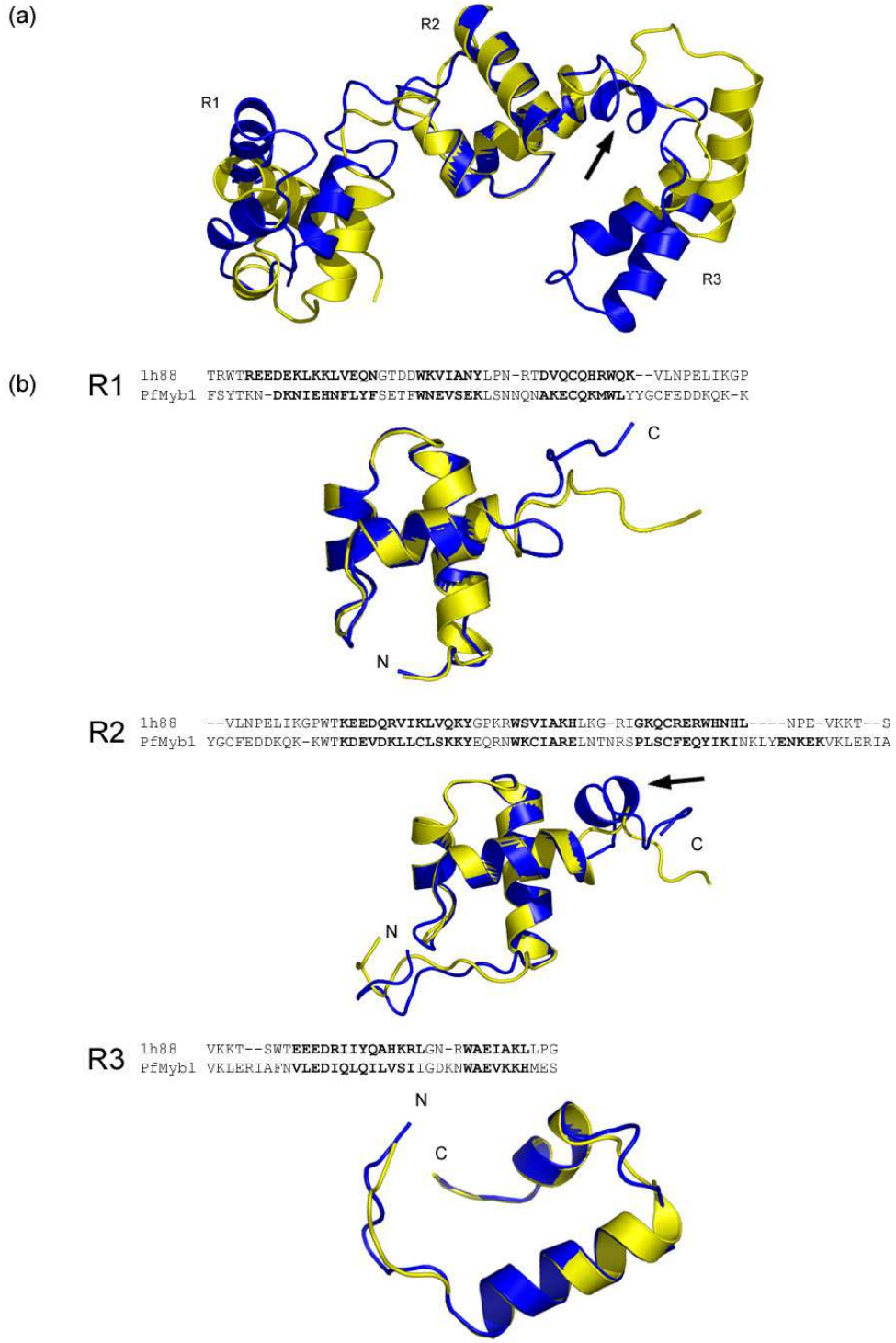


Figure 50. Domaine de liaison à l'ADN de PfMyb1 et répétitions R1, R2 et R3 superposés avec la structure support ayant servi à la modélisation.

La structure support 1H88 est colorée en jaune et le modèle 1 en bleu. Les structures ont été superposées grâce à l'option 'Iterative Magic Fit' de SwissPDB Viewer. **(a)** Domaine de liaison à l'ADN complet. La flèche indique l'hélice α supplémentaire. **(b)** Répétitions superposées à la structure support indépendamment les unes des autres. L'orientation de la chaîne peptidique est indiquée par N et C pour N-terminal et C-terminal. Au-dessus de chaque répétition se trouve l'alignement de séquences correspondant. Les résidus engagés dans les hélices sont en caractère gras.

Dans les répétitions R1 et R2 (Figure 50b), l'agencement des trois hélices est similaire. Les deux premières hélices sont quasiment parallèles et la troisième hélice forme un angle droit avec les deux autres. Pour la répétition R3, les deux premières hélices suivent le même arrangement que dans les répétitions R1 et R2, mais la troisième hélice n'ayant pas été modélisée, on ne peut pas connaître la trajectoire qu'elle aurait empruntée. On retrouve donc, pour R1 et R2, une architecture en « Orthogonal Bundle ».

Intéressons-nous à chacune des répétitions en détail. La première hélice de la répétition R1 commence plus tard par rapport à la structure support (Figure 50b). Les structures sont ensuite bien superposées jusqu'à la boucle reliant les deuxième et troisième hélices : celle-ci est plus longue d'un résidu dans la structure modèle (résidu en rouge sur la Figure 51a). La troisième hélice est plus courte d'un résidu (Figure 51a) ce qui fait que la boucle reliant la première et la deuxième répétition commence plus tôt (Figure 51b). Cette boucle a une longueur différente entre la structure modèle et la structure support et présente, de plus, une trajectoire différente : tout ceci est dû à la présence d'insertions-délétions dans l'alignement ayant servi à la modélisation, aussi bien dans la séquence cible que dans la séquence support (résidus en rouge sur la Figure 51b). Dans la répétition R2, les première et deuxième hélices ainsi que la boucle les reliant sont correctement superposées. Tout comme dans la répétition R1, pour la structure modèle, la boucle reliant les deuxième et troisième hélices est plus longue d'un résidu (résidu en rouge sur la Figure 51c) et la troisième hélice se termine un résidu plus tôt (Figure 51d). La boucle qui unit les répétitions R2 et R3 correspond à la partie de l'alignement contenant le plus d'insertions-délétions (résidus en rouge sur la Figure 51d) : le programme Modeller a donc modélisé une petite hélice de 5 résidus au début de cette boucle. Enfin, dans la répétition R3, la première hélice du modèle est plus courte ce qui fait que la boucle reliant les première et deuxième hélices est plus longue de deux résidus par rapport à la boucle de la structure support (Figure 51e).

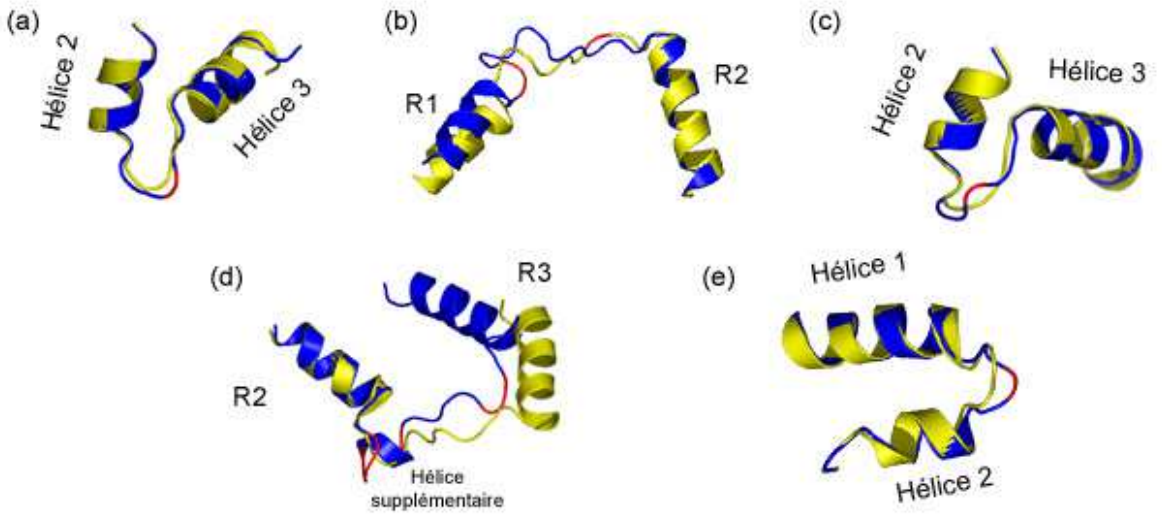


Figure 51. Détails du modèle 1 du domaine de liaison à l'ADN de PfMyb1.
Les structures support et modèle sont respectivement représentées en jaune et bleu. Les résidus colorés en rouge correspondent aux insertions dans l'une ou l'autre structure. **(a)** Hélices 2 et 3 de la répétition R1. **(b)** Boucle reliant les répétitions R1 et R2. **(c)** Hélices 2 et 3 de la répétition R2. **(d)** Boucle reliant les répétitions R2 et R3. **(e)** Hélices 1 et 2 de la répétition R3.

Passons maintenant à la deuxième méthode de modélisation par homologie, celle qui a consisté à modéliser le domaine de liaison à l'ADN en deux « morceaux » Les modèles retenus ont été appelés modèle 2.1 et modèle 2.2 pour la partie N-terminale et la partie C-terminale du domaine de liaison à l'ADN, respectivement.

Quand les structures modèles sont comparées aux structures support grâce au programme CE, les valeurs de RMSD sont bien meilleures que dans la première partie : de 3,3 Å, elles passent à 1,9 Å pour la partie N-terminale et à 2,6 Å pour la partie C-terminale. Les valeurs de Z-scores sont, quant à elles, toujours bonnes car supérieures à 4,5. Avec Verify3D et ProSa2003 (voir Figure 69, p. 242), la qualité des modèles n'est pas régulière ; en effet comme pour le modèle 1, la partie N-terminale du modèle 2.1 et la partie C-terminale du modèle 2.2 ne semblent pas très correctes d'un point de vue conformationnel. En revanche, la partie chevauchante obtient à peu près les mêmes scores dans les deux modèles. Chacun des deux modèles comporte deux résidus situés en dehors des zones autorisées du diagramme de Ramachandran (résidus surlignés en cyan sur la Figure 49) : ils correspondent tous à des glycines dans la structure ayant servi de support à la modélisation.

Si les structures modèles sont superposées à leur structure support en favorisant l'alignement structural de la répétition R2 du domaine de liaison à l'ADN de PfMyb1, les

autres répétitions, encore une fois, ne se superposent pas du tout avec la structure support, et ce dans le cas du modèle 2.1 (Figure 52a) comme dans celui du modèle 2.2 (Figure 52b).

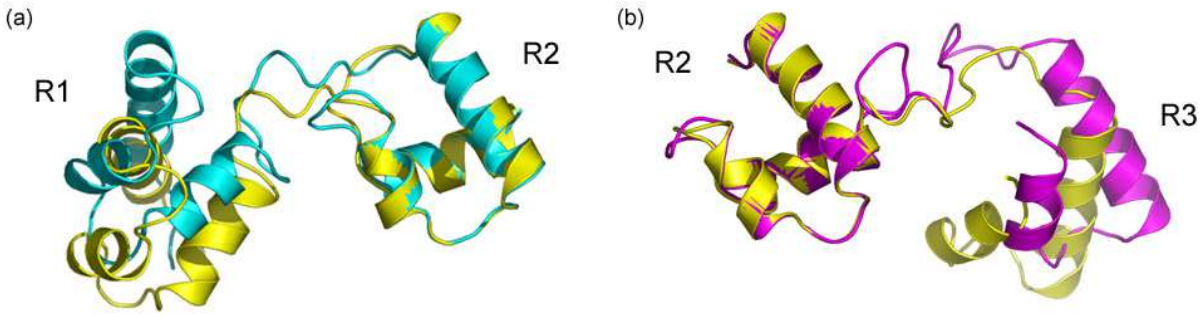


Figure 52. Modèles 2.1 et 2.2 du domaine de liaison à l'ADN à l'ADN de PfMyb1 avec la structure support ayant servi à la modélisation.
La structure support 1H88 est colorée en jaune, le modèle 2.1 en cyan **(a)** et le modèle 2.2 en magenta **(b)**. Les structures ont été superposées grâce au programme ProFit.

Cependant, lorsque les répétitions sont séparées artificiellement puis superposées indépendamment les unes des autres, les hélices modélisées se superposent très bien aux hélices ayant servi de support (Figure 53). Encore une fois, ce sont les boucles réunissant les différentes répétitions qui empruntent une trajectoire différente de celle empruntée dans la structure support ; la position relative des répétitions les unes par rapport aux autres s'en trouve donc perturbée.

Dans les répétitions R1 et R2 du modèle 2.1 ainsi que la répétition R2 du modèle 2.2 (Figure 53), les hélices adoptent la même architecture en « Orthogonal Bundle ». De plus, dans le modèle 2.2, aucune hélice α supplémentaire n'a été modélisée après la troisième hélice de la répétition R2, contrairement à ce qui a pu être observé dans le modèle 1.

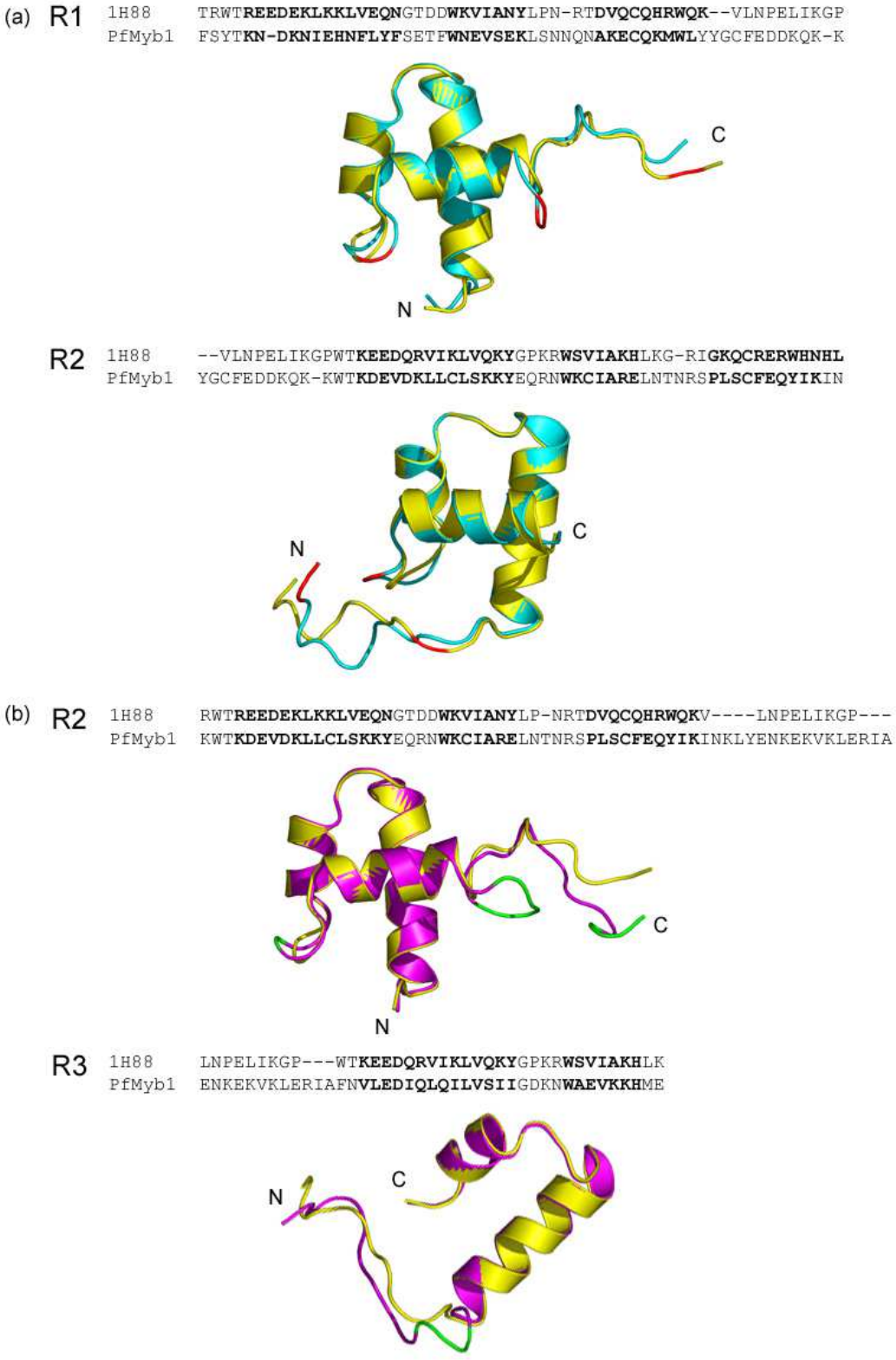


Figure 53. Répétitions des modèles 2.1 et 2.2 avec leur structure support propre.
La structure support 1H88 est colorée en jaune, les répétitions du modèle 2.1 en cyan **(a)** et les répétitions du modèle 2.2 en magenta **(b)**. Les résidus colorés en rouge (a) et en vert (b) correspondent aux insertions dans l'une ou l'autre structure. Les structures ont été superposées indépendamment les unes des autres grâce au programme ProFit. L'orientation de la chaîne peptidique est indiquée par N et C pour N-terminal et C-terminal. Au-dessus de chaque répétition se trouve l'alignement de séquences correspondant. Les résidus engagés dans les hélices sont en caractère gras.

Focalisons-nous tout d'abord sur le modèle 2.1. Dans la répétition R1 (Figure 53a-R1), on voit que la première hélice est plus courte d'un résidu à l'extrémité N-terminale à cause du gap introduit dans la séquence cible lorsqu'a été fait l'alignement de séquences nécessaire à la modélisation (Figure 49). Les structures sont ensuite très bien superposées jusqu'à la boucle reliant les deuxième et troisième hélices qui est plus longue d'un résidu dans la structure modèle (résidu en rouge sur la Figure 53a-R1). Enfin, la troisième hélice se termine un résidu plus tôt que dans la structure support. Ce qui fait que la boucle rejoignant la répétition R2 est au début un peu plus longue et forme une sorte de volute avant de reprendre une trajectoire similaire à celle de la boucle de la structure support ; ces deux boucles se séparent à partir du moment où il y a un résidu de plus dans la structure support. Mais le problème est inversé quand les répétitions R2 sont superposées (Figure 53a-R2) : alors que les parties N-terminales divergent, les parties C-terminales se rejoignent pour être finalement très bien superposées juste avant la première hélice de la répétition R2. Cette hélice ainsi que sa suivante et la boucle qui les relie sont très bien superposées avec la structure support alors que la boucle située entre la deuxième et la troisième hélices est elle aussi plus longue d'un résidu. Et pour finir, la troisième hélice se termine deux résidus avant celle de la structure support.

Passons maintenant au modèle 2.2. Dans la répétition R2 (Figure 53b-R2), les deux premières hélices et la boucle les reliant sont très bien superposées à la structure support. La boucle reliant les deuxième et troisième hélices est plus longue d'un résidu. La troisième hélice est très bien superposée à l'hélice de la structure support et a la même longueur. La boucle reliant les répétitions R2 et R3 est la boucle la plus difficile à modéliser car c'est celle qui contient le plus d'insertions-délétions dans l'alignement de séquences sur lequel est basée la modélisation (Figure 49b). Cette boucle a une longueur de 12 résidus dans la structure support alors que dans le modèle 2.2, elle comporte 19 acides aminés. C'est ce qui explique la présence de ce que l'on pourrait comparer à des épingles à cheveux : Modeller suit la structure support pour modéliser la cible et quand il est en présence de trop d'acides aminés, il fait des boucles dans la boucle. En revanche, les deux hélices de la répétition R3 (Figure 53b-R3) suivent très bien la structure support ainsi que la boucle qui les relie.

La deuxième méthode utilisée pour modéliser le domaine de liaison à l'ADN de PfMyb1 a consisté à modéliser ce domaine en deux parties qui seront ensuite réunies grâce à leur partie commune : la répétition R2. Il convient donc de comparer la répétition R2 du modèle 2.1 et la répétition R2 du modèle 2.2 (Figure 54). Quand les deux structures sont comparées, le programme CE calcule un RMSD de 0,9 Å ce qui montre que les deux structures sont très proches l'une de l'autre.

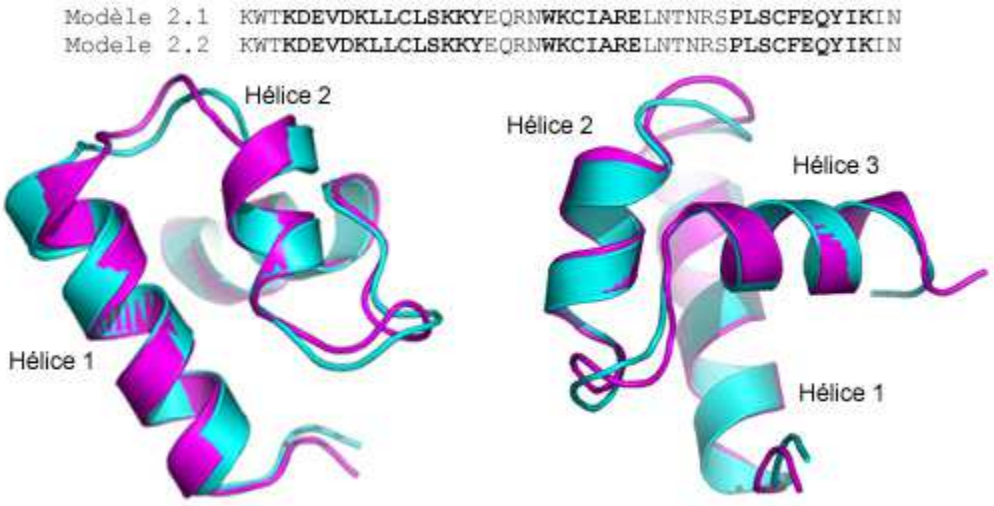


Figure 54. Superposition des répétitions R2 du modèle 2.1 et du modèle 2.2.
Le modèle 2.1 est coloré en cyan, le modèle 2.2 en magenta. Au-dessus de la superposition des deux structures se trouve l'alignement de séquences correspondant. Les résidus engagés dans les hélices sont en caractère gras.

En tout premier lieu, il convient de remarquer que la troisième hélice de la répétition R2 a la même longueur dans les deux modèles, alors que les hélices ayant servi de support pour la modéliser avaient des longueurs différentes (Figure 53). Si on regarde maintenant l'agencement dans l'espace de cette répétition, on voit que les deux premières hélices ne sont pas tout à fait orientées de la même manière bien que tout de même très proches. En fait, ce qui différencie principalement les deux modèles de la répétition R2, ce sont les boucles reliant les différentes hélices.

Dans toutes les structures de protéines Myb connues, à l'intérieur de chaque répétition, les « tryptophanes » sont orientés vers le cœur hydrophobe formé par les trois hélices, tout comme la cystéine des répétitions R1 et R2. La Figure 55 et la Figure 56 montrent la position

des chaînes latérales de ces différents acides aminés dans les répétitions des trois structures modèles obtenues.

En ce qui concerne la répétition R1 (Figure 55a), que ce soit dans le modèle 1 ou le modèle 2.1, les chaînes latérales des résidus Trp-22 et Trp-42 sont dans le même plan que les chaînes latérales des résidus correspondants dans la structure support et bien qu'orientés dans le même sens, les cycles sont légèrement décalés. En revanche, la tyrosine se situant à la place du premier tryptophane est orientée de manière totalement différente du tryptophane de la structure support, s'éloignant ainsi du cœur hydrophobe de cette répétition. Quant à la cystéine, la petite chaîne latérale reste dans le cœur hydrophobe même si son orientation est différente.

Pour ce qui est de la répétition R3 (Figure 55b), les modèles 1 et 2.2 ne présentent pas les mêmes résultats. En effet, le cycle de la phénylalanine 116 dans le modèle 1 est orientée dans le même sens que les cycles du tryptophane correspondant même s'il est situé dans un plan parallèle très proche, alors que dans le modèle 2.2, le cycle de cette même phénylalanine est perpendiculaire au plan formé par les cycles du tryptophane correspondant et se trouve ainsi plus éloigné de ce que l'on pourrait considérer comme le cœur hydrophobe de la répétition. A l'inverse, alors que les cycles du résidu Trp-136 du modèle 1 ne sont pas orientés comme ceux du tryptophane correspondant dans la structure support mais restent tournés vers ce que l'on pense être le cœur hydrophobe, le tryptophane 82 du modèle 2.2 est pratiquement confondu avec le tryptophane de la structure support.

Enfin, dans la répétition R2 (Figure 56), les chaînes latérales de deux tryptophanes, de la tyrosine et de la cystéine (Trp-56, Trp-76, Tyr-96 et Cys-92 pour les modèles 1 et 2.1 ; Trp-2, Trp-22, Tyr-42 et Cys-38 pour le modèle 2) sont relativement très proches des chaînes latérales des acides aminés qui leur ont servi de support, la plus grosse différence observée étant celle du Trp-56 du modèle 2.1. En revanche, dans la répétition R2 des modèles 2.1 et 2.2, la chaîne latérale de la tyrosine (en position 96 dans le modèle 2.1 et en position 42 dans le modèle 2.2), bien qu'orientée vers le cœur hydrophobe, ne se trouve pas dans le même plan dans les deux modèles alors que les chaînes latérales des deux tryptophanes et de la cystéine sont quasiment confondues.

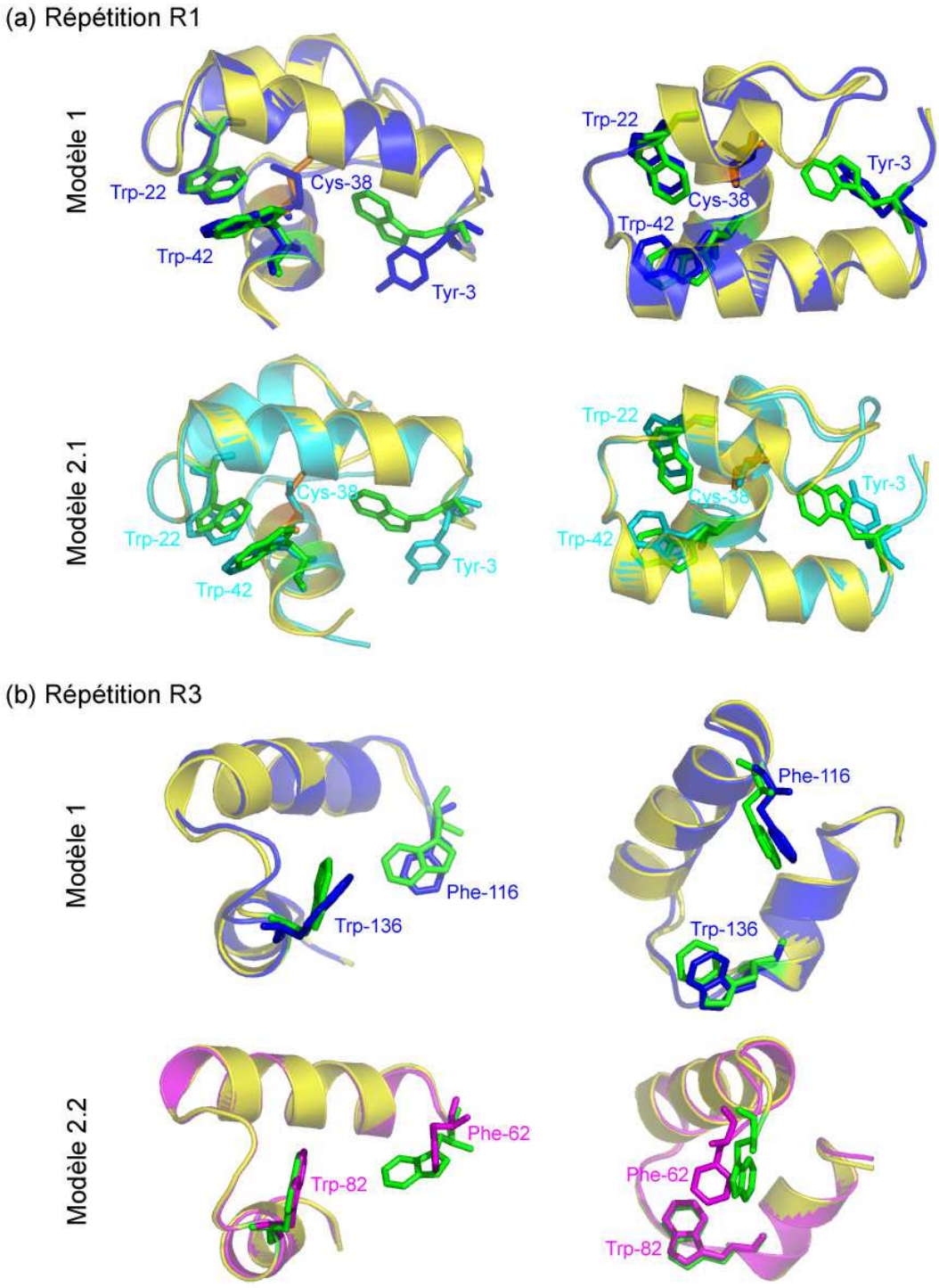


Figure 55. Orientation des chaînes latérales des répétitions R1 et R3.
(a) Répétition R1. **(b)** Répétition R3. Pour chaque répétition, deux vues différentes sont proposées. La structure support 1H88 est colorée en jaune, le modèle 1 en bleu, le modèle 2.1 en cyan et le modèle 2.2 en magenta. Les structures ont été superposées grâce à l'option 'Iterative Magic Fit' de SwissPDB Viewer ou avec le programme ProFit. Pour la structure support, en vert apparaissent les chaînes latérales des « tryptophanes » et en orange les chaînes latérales de la cystéine. Les chaînes latérales des résidus correspondants dans les structures modèles apparaissent dans la même couleur que le modèle et sont identifiées par la nature du résidu ainsi que sa position dans la séquence utilisée pour la modélisation (voir Figure 49).

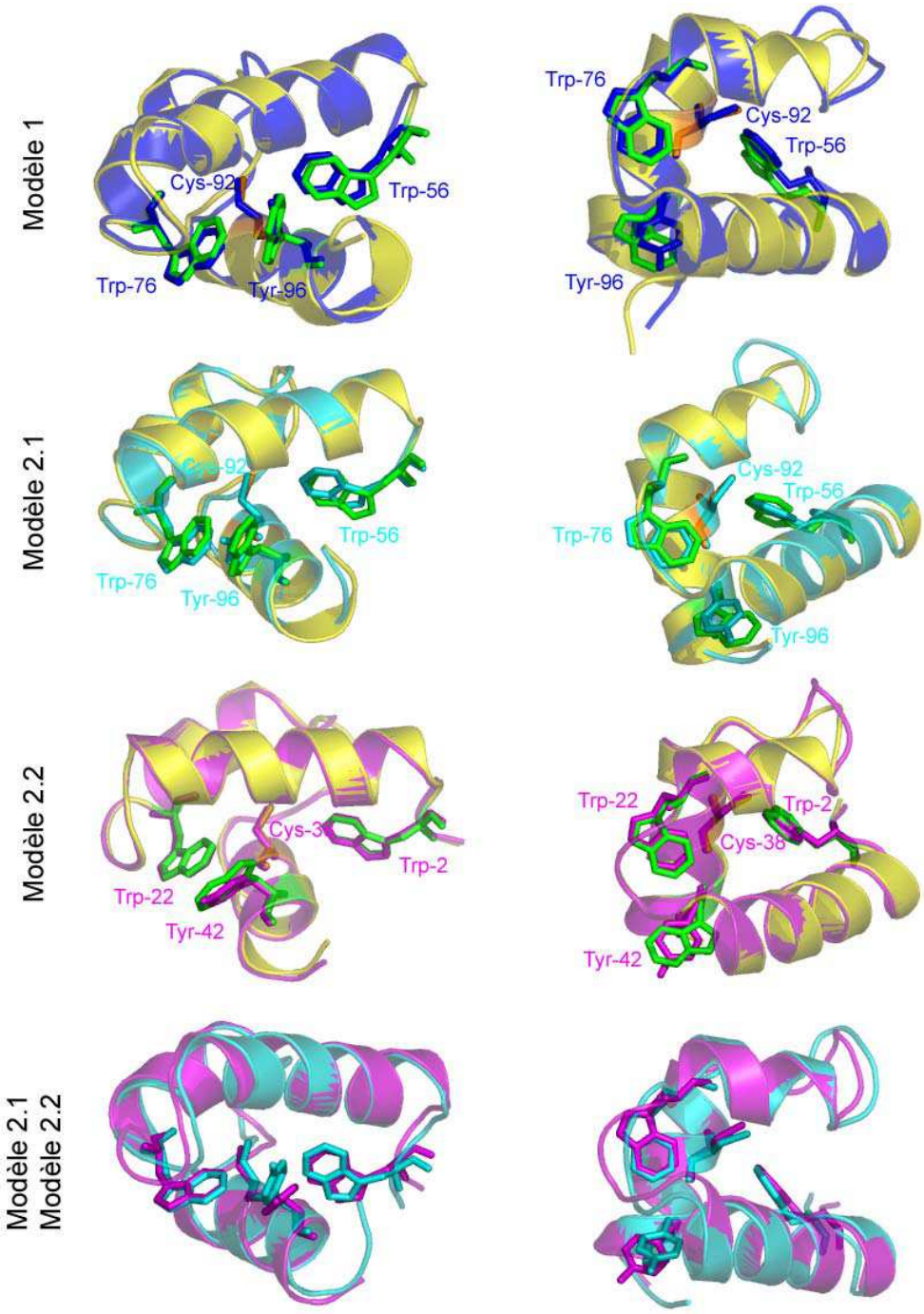


Figure 56. Orientation des chaînes latérales de la répétition R2.
La structure support est colorée en jaune, le modèle 1 en bleu, le modèle 2.1 en cyan et le modèle 2.2 en magenta. Les structures ont été superposées grâce à l'option 'Iterative Magic Fit' de SwissPDB Viewer ou avec le programme ProFit. Pour la structure support, en vert apparaissent les chaînes latérales des « tryptophanes » et en orange les chaînes latérales de la cystéine. Les chaînes latérales des résidus correspondants dans les structures modèles apparaissent dans la même couleur que le modèle et sont identifiées par la nature du résidu ainsi que sa position dans la séquence utilisée pour la modélisation (voir Figure 49).

La protéine c-Myb de souris utilisée pour ce travail de modélisation a été cristallisée en présence d'un double brin d'ADN. Les différentes structures modèles ont donc été visualisées avec ce double brin d'ADN. Comme le montre la Figure 57, si la répétition R2 de n'importe quel modèle semble être bien positionnée par rapport à l'ADN, ce n'est pas le cas des deux autres répétitions.

Pour le modèle 1 (Figure 57a), la répétition R1 est trop éloignée de la double hélice tandis que la répétition R3 est beaucoup trop proche de l'ADN et entre même en conflit avec celui-ci au niveau de la boucle située entre les deux hélices ainsi qu'au début de la deuxième hélice. De plus, quand l'alignement des structures se fait par rapport à la répétition R2, la répétition R3 passe du grand sillon au petit sillon de l'ADN. Quant aux chaînes latérales des acides aminés censés interagir avec l'ADN par homologie avec la structure support, les résultats ne sont pas très brillants. En effet, soit les chaînes latérales sont trop éloignées de l'ADN pour pouvoir interagir avec lui, soit elles sont en conflit avec celui-ci, soit elles sont bien positionnées par rapport aux chaînes latérales de la structure support mais trop courtes. Sur les 18 chaînes latérales analysées, seules sept d'entre elles semblent être bien positionnées pour pouvoir interagir avec l'ADN et sur ces sept chaînes latérales, six appartiennent à des acides aminés de la répétition R2 (Figure 58).

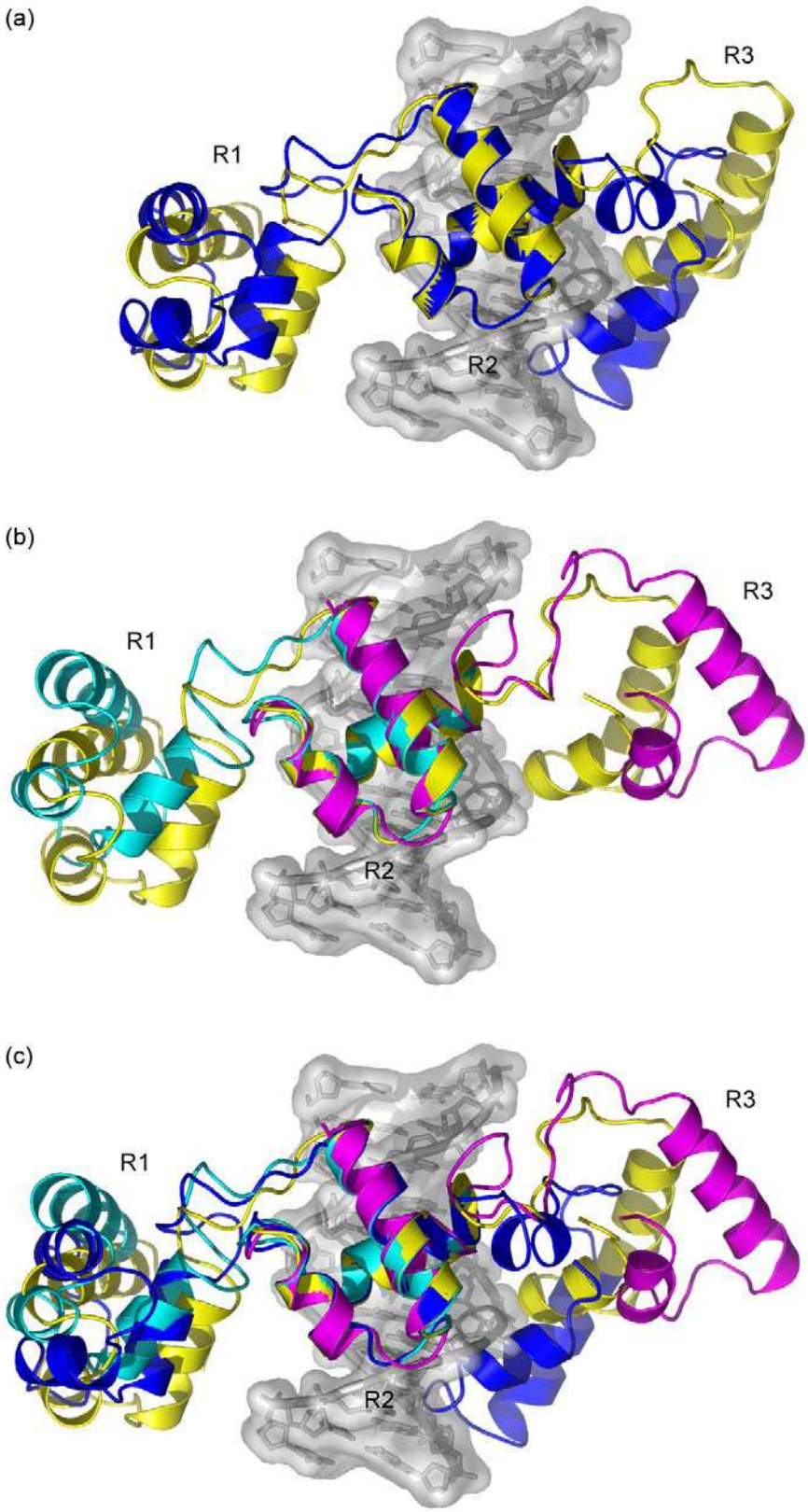


Figure 57. Visualisation des modèles de domaine de liaison à l'ADN de PfMyb1 avec la protéine c-Myb de souris ayant servi à la modélisation et un double brin d'ADN.
La structure support est colorée en jaune, le modèle 1 en bleu, le modèle 2.1 en cyan et le modèle 2.2 en magenta. **(a)** Modèle 1. **(b)** Modèles 2.1 et 2.2, superposés grâce au programme ProFit. **(c)** Modèle 1 et modèles 2.1 et 2.2.

Pour visualiser le modèle complet obtenu avec la deuxième méthode, les modèles 2.1 et 2.2 (Figure 57b) ont été réunis par l'intermédiaire de leur répétition R2 respective. Une fois encore, les répétitions R1 et R3 sont trop éloignées de l'ADN pour interagir avec lui quand la répétition R2 est correctement positionnée par rapport à l'ADN. Si on regarde la position des chaînes latérales des treize acides aminés de la répétition R2 du modèle 2.1 censés interagir avec l'ADN par homologie avec la structure support (Figure 58), il semble que sept d'entre elles soient correctement positionnées pour interagir avec l'ADN tandis que quatre autres sont « trop courtes » par rapport aux chaînes latérales de la structure support et deux autres trop éloignées. Les résultats sont donc assez similaires avec ceux obtenus pour la répétition R2 du modèle 1. Maintenant si on s'intéresse aux chaînes latérales des douze acides aminés de la répétition R2 du modèle 2.2 censés interagir avec l'ADN par homologie avec les acides aminés de la répétition R2 du modèle 2.1 (Figure 58), les résultats concordent avec ce que l'on a observé précédemment.

```
                K W                  KRWS             KQ RER H H
PfMyb1         QKKWTKDEVDKLLCLSKKYEQRNWKCIARELNTNRSPLSCFEQYIKIN
Modèle 1        c *                 ***c             -- **x - -
Modèle 2.1      x *                 x***             -- *** - -
Modèle 2.2        *                 *x*c             -- x** - x
```

Figure 58. États des chaînes latérales des acides aminés de la répétition R2 des différents modèles censés interagir avec l'ADN par homologie à la répétition R2 de la protéine c-Myb de souris.
Les résidus encadrés sont les résidus qui correspondent aux résidus en contact avec l'ADN dans la répétition R2 de la structure support, dont la nature est indiquée au dessus. Les chaînes latérales correctement positionnées sont indiquées par une astérisque (*), celles qui sont en conflit avec l'ADN par un « c », celles trop éloignées de l'ADN pour interagir avec lui par un « x » et les chaînes latérales bien positionnées par rapport aux chaînes latérales de la structure support mais trop courtes par rapport à celles-ci par un tiret (-).

IV.4 - Expériences biologiques

Les analyses *in silico* effectuées sur le facteur PfMyb1 et notamment sur son domaine de liaison à l'ADN nous font penser que le facteur PfMyb1 est un facteur de transcription capable de se fixer à l'ADN et de réguler la transcription de certains gènes. Parallèlement à ces études *in silico*, des expériences moléculaires *in vitro* ont donc été faites pour confirmer ces premiers résultats.

Pour résumer, la protéine PfMyb1 a été mise en évidence dans les extraits nucléaires de trophozoïtes grâce à une expérience de Western-blot couplée à une immunoprécipitation. De plus, des expériences de retardement sur gel ont montré que cette protéine était capable de se fixer sur des éléments de régulation de type *myb* : un élément de régulation issu du promoteur du gène *mim-1* de poulet qui a déjà été montré comme permettant la fixation d'une protéine de type Myb [277] et deux éléments de régulation identifiés *in silico* dans les promoteurs des gènes plasmodiaux *pfcrk1* [91] et *pfmap1* [92]. La spécificité de l'interaction a été vérifiée par retardement sur gel en présence d'anticorps spécifique et de compétiteurs spécifiques et aspécifiques. La fonction et les gènes cibles de ce facteur ont été analysés par interférence par ARN double brin sur des trophozoïtes, stade où l'expression du transcrit est la plus élevée, pour montrer les conséquences de la diminution de l'ARNm et donc de la protéine PfMyb1 sur la croissance et la régulation transcriptionnelle des gènes du parasite. Tout d'abord, la croissance est diminuée de 40% et la mort cellulaire a lieu au moment du passage de l'état trophozoïte à l'état schizonte. Cet état de fait ajouté à la diminution observée des niveaux d'expression du transcrit et de la protéine PfMyb1 au stade trophozoïte nous amènent à penser que le facteur de transcription PfMyb1 a un rôle essentiel dans la phase érythrocytaire du développement du parasite. De plus, des expériences effectuées avec la puce à ADN ciblée fabriquée dans notre laboratoire et composée de 153 gènes impliqués dans la régulation des gènes au sens large couplées à des expériences de RT-qPCR ont permis de mettre en évidence une différence d'expression significative pour huit gènes entre une culture témoin et une culture dont le niveau d'expression de *pfmyb1* est diminuée : un est surexprimé tandis que les sept autres sont sous-exprimés.

Les expériences biologiques menées sur ce facteur de transcription sont détaillées dans les articles suivants, situés à la fin de ce manuscrit :

- *Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 Plasmodium falciparum clones* (Mol Biochem Parasitol, 2004),
- *PfMyb1, a Plasmodium falciparum transcription factor, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation* (J Mol Biol, 2005).

IV.5 - Discussion et perspectives

Les protéines de type Myb sont très conservées au cours de l'évolution et caractérisées par leur domaine de liaison à l'ADN composé de deux ou trois répétitions en tandem (Figure 45). Grâce à ce domaine de liaison à l'ADN, une protéine appelée PfMyb1 a été annotée sur le chromosome 13 de *Plasmodium falciparum*. Alors que dans la plupart des protéines Myb, ce domaine de liaison à l'ADN est situé en N-terminal de la protéine, le domaine de liaison à l'ADN de la protéine PfMyb1 se trouve en C-terminal ; mais ceci a déjà été observé dans trois protéines Myb de *Dictyostelium discoideum :* DdMybB [292], DdMybC [155] et DdMybD [101], la fonction des deux premières ayant déjà été démontrées (Tableau 14).

Les répétitions composant le domaine de liaison à l'ADN des protéines Myb, aussi appelées domaines Myb, sont constituées d'une cinquantaine d'acides aminés dont trois tryptophanes régulièrement séparés par 18 ou 19 acides aminés. Les tryptophanes peuvent être remplacés par les deux autres acides aminés aromatiques : la phénylalanine et la tyrosine, ou encore par une isoleucine [235, 251]. Alors que les deux premières répétitions (R1 et R2) de la protéine PfMyb1 sont fidèles à ce schéma, la troisième répétition (R3) possède une insertion de 6 acides aminés entre le tryptophane central et la phénylalanine de la fin de la répétition (Figure 46). Une insertion de même longueur est observée dans la troisième répétition des protéines DdMybB, DdMybC, DdMybD et DdMybE, et il existe des protéines de type Myb connues pour être fonctionnelles qui contiennent des insertions beaucoup plus longues (Figure 47) : 23 acides aminés dans la répétition R1 de la protéine BAS1 de *Saccharomyces cerevisiae* [174, 377] et 54 acides aminés dans la répétition R3 de la protéine REB1 de *Kluyveromyces lactis* [264]. De plus, la présence d'insertions a aussi été observée dans de nombreuses protéines de *Plasmodium* [306], notamment dans les ARN polymérases I, II et III [123, 231, 232]. Il existe, en plus des « tryptophanes », d'autres résidus conservés dans chacune des répétitions (Figure 47) : ces résidus ne sont pas tous retrouvés dans PfMyb1, comme dans les protéines de *D. discoideum* ou des levures *S. cerevisiae*, *S. pombe* et *K. lactis*. Néanmoins, les résidus qui ont une fonction connue sont présents, comme la cystéine située entre les deuxième et troisième « tryptophanes » dans les répétitions R1 et R2. La réduction de la cystéine de la répétition R2 est indispensable à la bonne conformation de

la répétition R2, ce qui fait que la liaison de la protéine Myb à l'ADN *in vitro* et *in vivo* est dépendante des conditions d'oxydoréduction de la protéine [147, 153, 268].

Il est aussi intéressant de noter que les domaines Myb, qui sont présents en plusieurs exemplaires dans une même protéine, se répartissent en trois familles distinctes : R1, R2 et R3 (Figure 47) ; chaque domaine est plus proche des autres domaines de sa famille que des autres domaines de la même protéine, ce qui signifie que la duplication qui a généré ces répétitions a eu lieu avant la divergence des espèces [235].

La modélisation « tout automatique » de la protéine PfMyb1 a posé plus de problèmes que celle des protéines PfHMGB1 & PfHMGB2. Les résultats obtenus avec le méta-serveur @TOME présentés dans ce manuscrit et ceux présentés dans l'article *Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 Plasmodium falciparum clones* (Mol Biochem Parasitol, 2004) ne concordent pas. La modélisation présentée dans l'article a été faite en 2003, la requête a donc été relancée en 2005 pour confirmer les résultats ou avoir une meilleure structure support. De nouvelles structures sont apparues dans la PDB et les résultats de la modélisation en ont été chamboulés. A la structure considérée par TITO comme étant le meilleur support pour modéliser le domaine de liaison à l'ADN de PfMyb1 sont associés un pourcentage d'identité, une longueur d'alignement et une E-value qui en font, à mon avis, un très mauvais support. C'est pourquoi, je ne suis pas allée plus loin dans cette nouvelle modélisation avec @TOME et TITO. Il est décrit, dans l'article *Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 Plasmodium falciparum clones* (Mol Biochem Parasitol, 2004), une structure modèle avec trois hélices dans la répétition R2 et deux dans la répétition R3 (Figure 59). Elle a été modélisée par homologie avec la répétition R1 et le début de la répétition R2 de la structure de la protéine c-Myb de souris (fichier PDB : 1H88) et on peut voir que la répétition R2 est conforme à l'architecture « Orthogonal Bundle ». Cependant, comme dans le cas des facteurs PfHMGB, les insertions qui existaient dans l'alignement de la séquence support et de la séquence cible n'ont pas été modélisées.

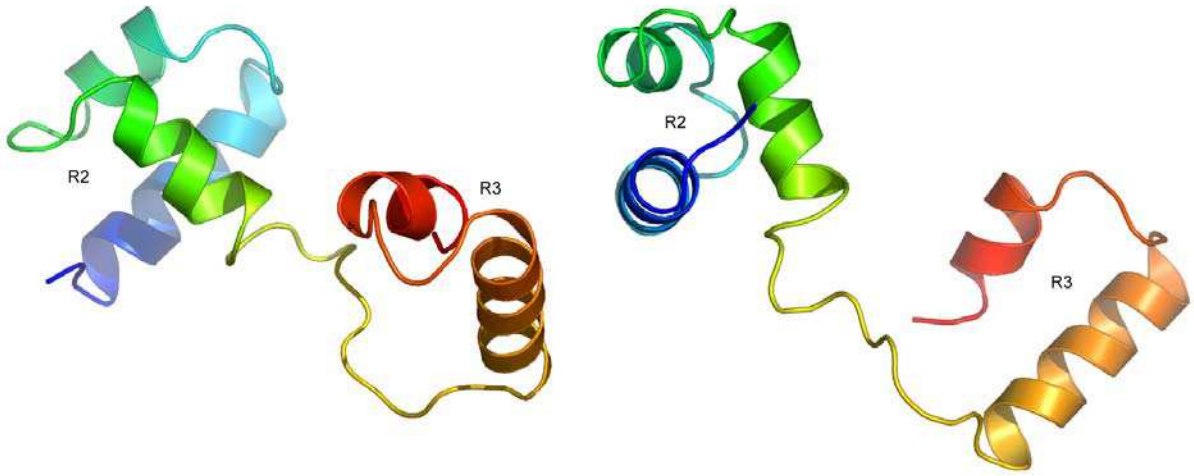


Figure 59. Modèle d'une partie du domaine de liaison à l'ADN de PfMyb1 obtenu par homologie avec le facteur c-Myb de souris.
La structure est visualisée sous deux angles différents et est colorée selon le spectre de la lumière, la partie N-terminale de la structure étant en bleu et la partie C-terminale en rouge.

Pour la deuxième méthode de modélisation, la structure de la protéine c-Myb de souris (fichier PDB 1H88) [365] a été utilisée comme support pour modéliser le domaine de liaison à l'ADN de PfMyb1. Il serait logique de se dire que la E-value associée à la structure 1H88 est trop élevée pour que cette structure puisse être utilisée comme support. Mais les structures de domaine de liaison à l'ADN que l'on trouve dans la PDB proviennent toutes de la protéine c-Myb de souris, de la protéine B-Myb de poulet ou encore de la protéine v-Myb du virus de la myéloblastose aviaire qui est quasiment identique à sa contrepartie cellulaire. Aucune protéine Myb provenant d'un eucaryote inférieur, ou son domaine de liaison à l'ADN, n'a été cristallisée et comme l'alignement des différentes répétitions a pu le montrer (Figure 47), les domaines Myb de PfMyb1, comme les domaines Myb des autres eucaryotes inférieurs, divergent un peu par rapport aux domaines Myb des eucaryotes supérieurs. Cependant, j'ai considéré que cette structure était valable en me basant aussi sur les résultats obtenus avec le méta-serveur @TOME : à chaque fois que la structure 1H88 a été sélectionnée par un programme, elle était considérée comme étant un résultat significatif, contrairement à la structure désignée par TITO comme étant le meilleur support (données non montrées).

La comparaison du Tableau 16 et de la Figure 48, montre qu'il existe une différence de longueur d'alignement et de pourcentage d'identité : parfois la séquence utilisée pour faire l'alignement quand la base de données PDB est interrogée par BlastP ne correspond pas à ce qui a vraiment été cristallisé. Dans notre cas, la séquence présente dans le champ SEQRES du fichier PDB et qui a été utilisée par BlastP pour faire l'alignement correspond aux résidus 35

à 193 de la séquence c-Myb de souris, alors que les coordonnées atomiques du fichier PDB correspondent aux acides aminés de la position 39 à la position 190.

Le domaine de liaison à l'ADN de PfMyb1 a été modélisé de deux manières (Figure 48 & Figure 49). Dans les deux cas, la fin de la répétition R3 n'a pas été inclue dans les alignements car cela aurait fait chuter le pourcentage d'identité qui est déjà faible pour une modélisation par homologie (limite : 25%).

Dans les trois modèles obtenus, certains résidus sortent des zones autorisées du diagramme de Ramachandran (résidus surlignées en cyan sur la Figure 49). Ils correspondent tous, sauf le troisième résidu du modèle 1, à des glycines de la séquence support qui, elles aussi, sortent des zones autorisées du diagramme de Ramachandran. La glycine est un acide aminé particulier car, comme sa chaîne latérale se résume à un atome d'hydrogène, cela lui confère une plus grande flexibilité. Il existe d'ailleurs un diagramme de Ramachandran dédié à la glycine et un autre à la proline, qui est aussi un acide aminé particulier, vu que la chaîne latérale est rattachée à la chaîne principale [167].

Les structures modèles et support ont toutes été superposées en favorisant la superposition de la répétition R2 du modèle avec la répétition lui ayant servi de support car c'est la répétition la mieux conservée : c'est la seule répétition identifiée comme étant un domaine Myb par le programme MotifScan et quand on analyse la qualité des trois modèles avec Verify3D et ProSa2003, c'est la partie de la structure qui obtient les meilleurs scores. Dans le cas des modèles 1 et 2.1, cette superposition s'est faite correctement avec le programme SwissPDB Viewer, alors que dans le cas du modèle 2.2, l'utilisation du programme ProFit a été nécessaire pour imposer les segments de structures à superposer.

Le modèle 1 (Figure 50a), composé des trois hélices dans les répétitions R1 et R2 et de deux hélices dans la répétition R3, comporte une petite hélice supplémentaire située juste après la troisième hélice de la répétition R2. Le modèle 2.1 est composé de trois hélices pour les répétitions R1 et R2 tandis que le modèle 2.2 comporte trois hélices dans la répétition R2 et deux hélices dans la répétition R3 (Figure 52). Contrairement au modèle 1, aucune nouvelle hélice n'a été modélisée après la répétition R2 dans le modèle 2.2. Quand on superpose la structure modèle à sa structure support, il apparaît que, dans chaque structure

modèle, les domaines Myb ne sont pas positionnés de la même manière les uns par rapport aux autres dans l'espace : les répétitions R1 et R3 ne sont donc pas bien superposées aux répétitions qui leur ont servi de support, ce qui explique les valeurs de RMSD moyennement satisfaisantes calculées par le programme CE. Néanmoins, si les répétitions sont considérées indépendamment les unes des autres dans chaque modèle (Figure 50b et Figure 53), les hélices de chaque répétition se superposent avec le support de manière très correcte dans l'ensemble. Cependant, certaines hélices se terminent un ou deux résidus plus tôt dans la structure modèle par rapport à la structure support, comme la troisième hélice des répétitions R1 et R2 des modèles 1 et 2.1 ou la première hélice de la répétition R3 du modèle 1. Et dans les modèles 1 et 2.1, la première hélice de la répétition R1 commence plus tard par rapport à l'hélice qui lui a servi de support à cause du gap présent dans la séquence cible (Figure 50b et Figure 53a). Mais l'arrangement de ces hélices les unes par rapport aux autres dans chaque répétition R1 ou R2 est conforme à l'architecture appelée « Orthogonal Bundle » de la base de données CATH, à savoir deux hélices antiparallèles avec une troisième hélice perpendiculaire aux deux premières, formant comme une lettre H (Figure 50b et Figure 53). De plus, il est très intéressant de noter que la répétition R2 du modèle 2.1 et la répétition R2 du modèle 2.2 sont quasiment similaires alors qu'elles ont été modélisées pour la première à partir de la répétition R2 et pour la deuxième à partir de la répétition R1 de la protéine c-Myb de souris (Figure 54).

Dans chacune des répétitions, les chaînes latérales des trois « tryptophanes » sont orientées vers le cœur hydrophobe formé par les trois hélices [235], de même que la cystéine des répétitions R1 et R2. Dans l'ensemble, les chaînes latérales des structures modèles sont orientées comme les chaînes latérales des structures support (Figure 55 et Figure 56), mis à part la tyrosine 3 de la répétition R1 des modèles 1 et 2.1 et la phénylalanine 62 de la répétition R3 du modèle 2.2, même si, dans ce dernier cas, on ne peut qu'extrapoler l'emplacement de la troisième hélice et donc du cœur hydrophobe en observant les répétitions R1 et R2. Si les chaînes latérales sont analysées par le programme SCWRL, la majorité semble être dans une bonne conformation. Le programme a repositionné certaines chaînes latérales ce qui fait que ces résidus ne sont plus orientés vers le cœur hydrophobe des répétitions (Figure 60) :

- dans la répétition R1, la chaîne latérale du deuxième tryptophane (Trp-22) de manière quasiment identique dans les modèles 1 et 2.1,
- dans la répétition R2, le premier tryptophane du modèle 1 (Trp-56) et le deuxième tryptophane dans le modèle 2.2 (Trp-22),
- dans la répétition R3, le deuxième tryptophane des modèles 1 (Trp-136) et 2.2 (Trp-82).

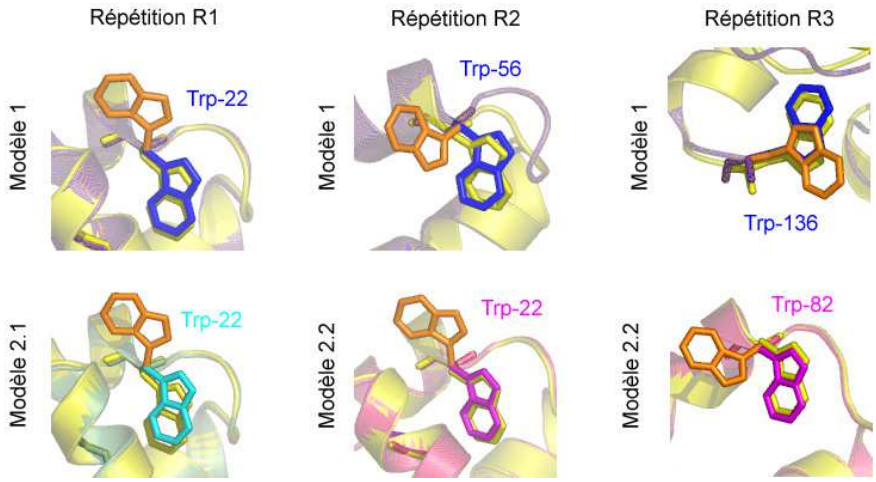


Figure 60. Chaînes latérales repositionnées par le programme SCWRL.
La structure support est colorée en jaune, le modèle 1 en bleu, le modèle 2.1 en cyan et le modèle 2.2 en magenta. Les chaînes latérales repositionnées par SCWRL apparaissent en orange.

En plus des résidus caractéristiques que sont les « tryptophanes » et la cystéine, trois acides aminés de la répétition R2 ont été montrés comme étant importants pour l'activation de la transcription du gène *mim-1* de poulet , une des cibles connues de la protéine c-Myb [180, 277]. Il s'agit des résidus Ile-91, Leu-106 et Val-117. Le promoteur de ce gène présente des sites de liaison pour les protéines Myb et C/EBPβ ; l'activation synergique du gène *mim-1* par ces deux facteurs de transcription suggère que ces deux protéines sont en contact direct [192], soit grâce au faible espace entre les deux sites de liaison [46, 276], soit parce que l'ADN fait une boucle qui permet de rapprocher les deux facteurs l'un de l'autre alors que leurs sites de liaison sont distants [366]. La structure du complexe R2R3-ADN publiée en 1992 (fichiers PDB : 1MSE/1MSF) [284, 285] indique que ces trois résidus hydrophobes sont exposés au solvant et donc disponibles pour une interaction protéine-protéine avec C/EBPβ. Dans le cas de la structure support utilisée ici et des modèles obtenus, les chaînes latérales de

ces résidus sont aussi exposées au solvant (Figure 61) et pourraient donc interagir avec une protéine C/EBPβ présente dans le noyau du parasite.

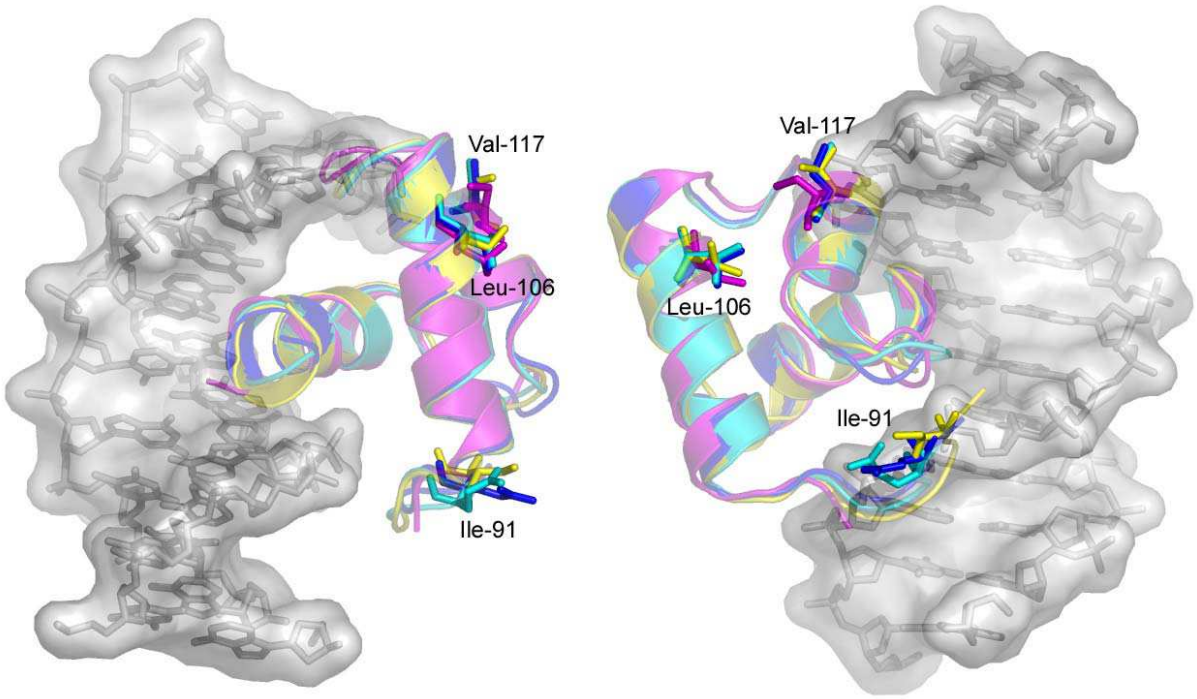


Figure 61. Trois résidus hydrophobes de la répétition R2 exposés au solvant.
La structure support est colorée en jaune, le modèle 1 en bleu, le modèle 2.2 en cyan et le modèle 2.2 en magenta. La numérotation des résidus correspond à la structure support.

Après avoir vu les résidus de PfMyb1 qui pourraient intervenir dans les interactions de notre facteur de transcription avec d'autres protéines, intéressons-nous à la structure générale du domaine de liaison à l'ADN de PfMyb1. Les domaines Myb de chaque modèle ne sont pas positionnés dans l'espace les uns par rapport aux autres comme dans la structure modèle. Ceci est dû aux boucles reliant les différents domaines qui n'ont pas la même longueur dans les structures modèles et dans les structures support (Figure 51 et Figure 53). Effectivement, mis à part le gap introduit dans la première hélice de la répétition R1 de PfMyb1, tous les autres ont été introduits entre les domaines conservés, que ce soit dans la séquence cible ou dans la séquence support (Figure 49). L'introduction de ces gaps rend la modélisation plus difficile pour le programme Modeller et comme aucune indication n'est donnée sur la position relative des domaines, les boucles sont modélisées sans aucune contrainte. De ce fait, dans les trois modèles, alors que la répétition R2 est correctement positionnée par rapport au grand sillon de l'ADN, les répétitions R1 et R3 ne le sont pas du

tout (Figure 57). Dans les modèles 1 et 2.1, la répétition R1 est plus éloignée de l'ADN que dans la structure support ; quant à la répétition R3, dans le cas du modèle 1, elle est passé dans le petit sillon et entre en conflit avec l'ADN alors que dans le cas du modèle 2.2, elle est beaucoup trop éloignée pour pouvoir interagir avec lui. Le modèle 1 a été superposé à la structure support en favorisant soit la répétition R1 soit la répétition R3. Dans le premier cas (Figure 62a), les répétitions R2 et R3 se retrouvent beaucoup trop éloignées de l'ADN alors que ce sont les deux répétitions qui sont responsables de la liaison à l'ADN ; dans le deuxième cas (Figure 62b), la répétition R1 se retrouve plus proche du petit sillon que du grand sillon de l'ADN et la répétition R2 se retrouve souvent dans l'ADN.

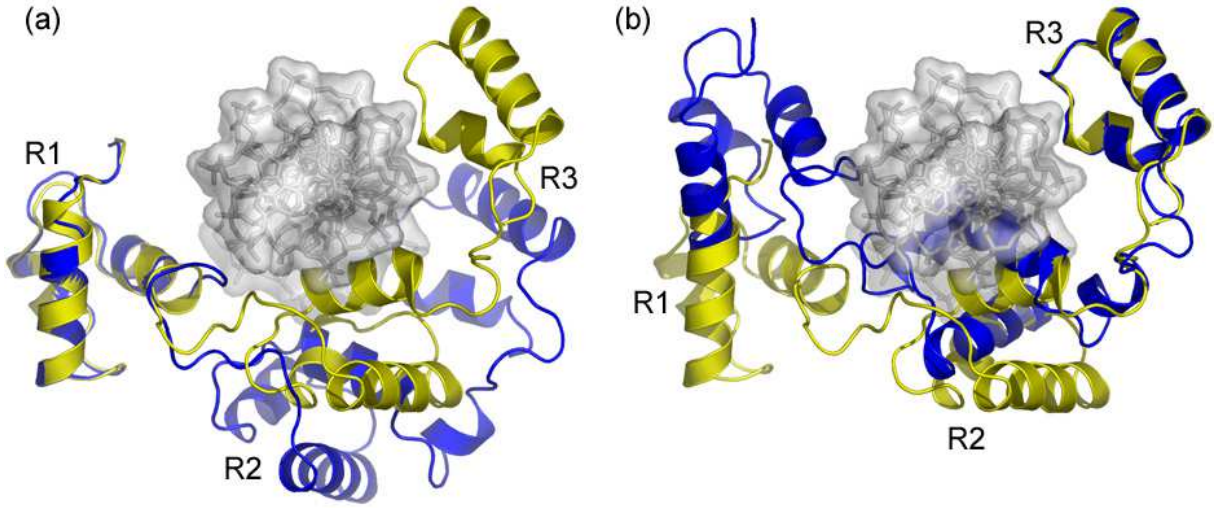


Figure 62. Superposition de la structure modèle et de la structure support.
La structure support est colorée en jaune et le modèle 1 en bleu. Les structures ont été superposées grâce au programme ProFit. **(a)** Superposition des deux structures par rapport à R1. **(b)** Superposition des deux structures par rapport à R3.

Les cinq meilleures structures modèles obtenues avec Modeller lors de la modélisation avec un seul alignement (à savoir le modèle 1 de mon étude et les quatre structures modèles ayant obtenu les meilleures fonctions objectives après le modèle 1) ont été comparées (Figure 63). Les quatre modèles supplémentaires présentent les mêmes problèmes que le modèle 1 : quand la répétition R2 est « correctement » positionnée par rapport à l'ADN, les deux autres répétitions ne se positionnent pas de façon satisfaisante pour que les structures soient de bons modèles dans leur intégralité. Il est aussi intéressant de noter que sur les cinq meilleures structures de cette simulation, seules deux ont une hélice supplémentaire après la répétition R2 : le modèle 1 et la structure n° 71 (indiquée par une flèche).

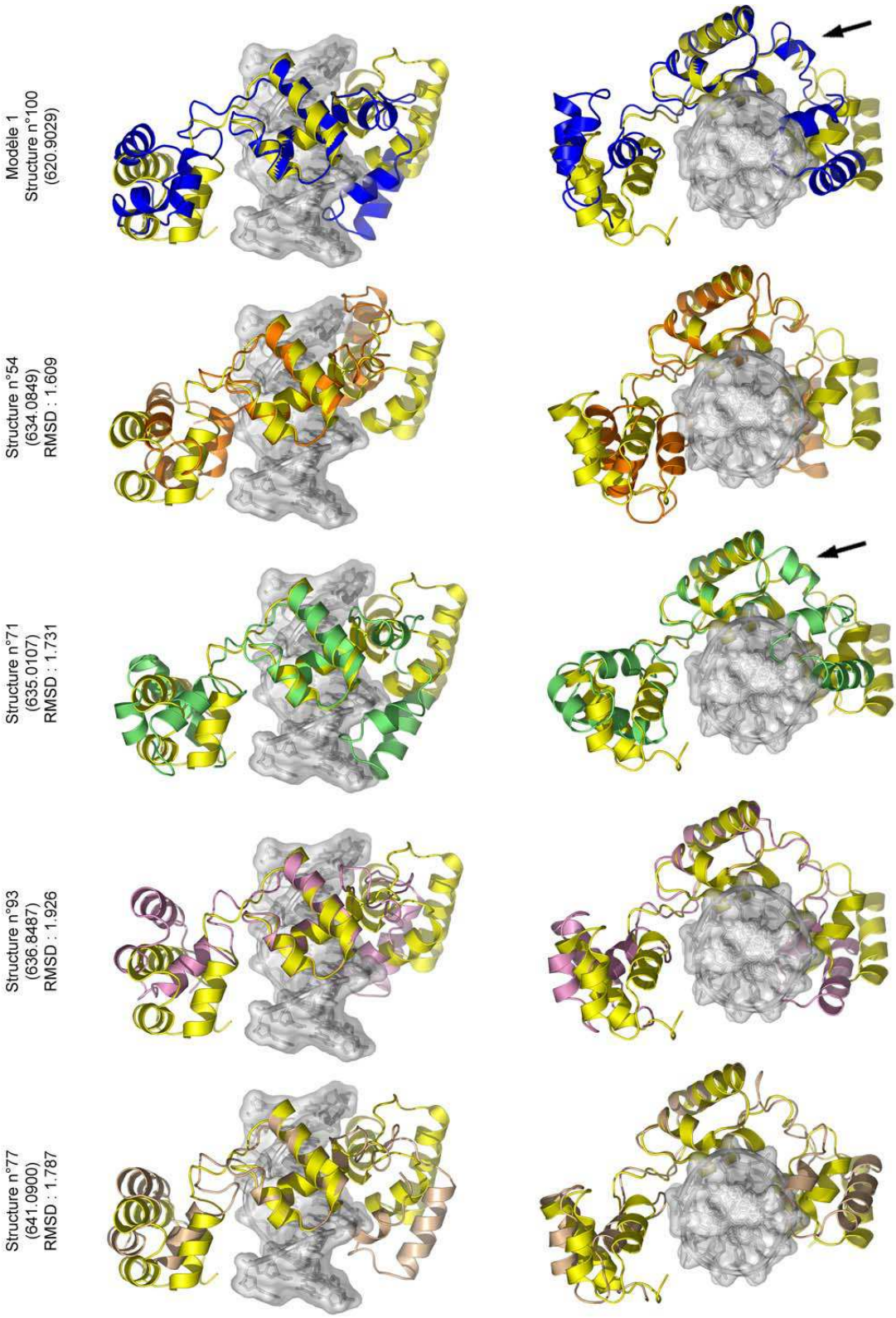


Figure 63. Les quatre meilleures structures modèles obtenues après le modèle 1.
La structure support 1H88 est colorée en jaune et le modèle 1 en bleu. Les autres structures modèles obtenues avec Modeller sont présentées dans l'ordre de leur fonction objective indiquée entre parenthèses, avec la valeur de RMSD calculée par le programme ProFit lorsque ces structures ont été superposées au modèle 1, en privilégiant la répétition R2. La flèche indique l'hélice α supplémentaire entre R2 et R3.

Il est difficile de discuter sur la position des chaînes latérales des résidus appartenant aux répétitions R1 et R3 de PfMyb1 qui sont censés interagir avec l'ADN, par homologie à la protéine c-Myb de souris, car ces domaines Myb sont très mal placés par rapport à l'ADN et donc les chaînes latérales seront forcément mal positionnées par rapport à l'ADN même si elles sont correctes d'un point de vue conformationnel. En ce qui concerne la répétition R2, treize chaînes latérales sont à prendre en considération (Figure 58). Il existe trois zones principales d'interaction avec l'ADN dans la répétition R2. Dans la première zone, la lysine entre en conflit avec l'ADN dans le modèle 1 et semble en être trop éloignée dans le modèle 2.1 ; la lysine est en fait un acide aminé dont la chaîne latérale assez longue présente quatre angles χ et est donc capable d'adopter un certain nombre de conformations, dont une pourra correspondre à ce qui se passe pour la lysine de la structure support. Dans la deuxième zone d'interaction, chaque chaîne latérale est au moins une fois correctement positionnée pour interagir avec l'ADN. Quant à la troisième zone d'interaction, elle comporte un plus grand nombre de résidus différents. Il semblerait donc logique que cette dernière zone subisse une réorganisation pour que l'interaction avec l'ADN se fasse correctement. En effet, certaines chaînes latérales sont plus courtes dans la structure modèle que dans la structure support (sérine à la place de glutamine ou phénylalanine à la place d'arginine). De plus, les résidus lysine et histidine, qui sont des résidus chargés positivement interagissant facilement avec l'ADN chargé négativement, sont remplacés par des résidus plutôt hydrophobes (leucine et isoleucine), ce qui laisse penser qu'ils ne seront plus impliqués dans l'interaction avec l'ADN et remplacés par d'autres résidus.

Il a été démontré, il y a 15 ans, que les protéines c-Myb, ainsi que les protéines v-Myb, reconnaissent dans l'ADN le motif AAC[acgt]G[acgt][act] [31, 270]. Quelques années plus tard, Jun Tanikawa et ses collaborateurs ont montré que c-Myb interagit spécifiquement avec les première, troisième et cinquième bases du motif et plus particulièrement que les répétitions R3 et R2 se placent respectivement en face des parties AAC et G[acgt][act] dans le grand sillon de l'ADN [371]. Et en 1996, Chie Kanei-Ishii et ses collaborateurs précisent que les trois nucléotides-clés de la séquence ADN sont liés respectivement aux résidus Asn-183, Lys-182 et Lys-128 [192]. Parmi les quatre fichiers PDB qui comportent un domaine de liaison à l'ADN complet ou partiel d'une protéine Myb lié à un double brin d'ADN (Tableau 16),

deux correspondent aux données expérimentales de C. Kanei-Ishii et les deux autres s'en rapprochent. Dans le domaine de liaison à l'ADN de la protéine c-Myb de souris qui a servi à la modélisation (fichier PDB 1H88), le résidu Lys-182 ne se fixe pas sur la cytosine en troisième position mais sur le nucléotide précédent alors que les résidus Asn-183 et Lys-128 se fixent « comme il faut » sur la cytosine en première position et la guanine en cinquième position. Mais c'est le seul fichier de la base de données PDB dans lequel se trouve la structure du domaine de liaison à l'ADN complet des protéines Myb comportant trois répétitions. Et ce domaine de liaison à l'ADN est complexé à un double brin d'ADN (correspondant à une partie du promoteur du gène *tom-1*) sur lequel est aussi fixé la protéine C/EBPβ humaine [365]. L'interaction entre c-Myb et C/EBPβ peut entraîner des contraintes et ainsi avoir une incidence sur la liaison des deux protéines sur le double brin d'ADN.

Le facteur PfMyb1 se fixe de manière spécifique à l'ADN sur une séquence répondant au consensus cité précédemment (voir *Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 Plasmodium falciparum clones*, 2004) mais il n'existe aujourd'hui aucune donnée expérimentale qui donnerait des indications sur la structure réelle et/ou sur les résidus impliquées dans la liaison à l'ADN. Il y a donc deux hypothèses à considérer.

Tout d'abord, si on considère que la structure support choisie pour modéliser le domaine de liaison à l'ADN de PfMyb1 est la bonne, les modèles générés lors de ce travail peuvent donc être considérés comme un bon début car les carbones α des hélices semblent être positionnés correctement dans chacune des trois répétitions, prises indépendamment les unes des autres. Néanmoins, la méthode de modélisation *ab initio* sera nécessaire pour modéliser la fin de la répétition R3 ainsi que les boucles reliant ces différentes répétitions tout en introduisant des contraintes sur la position relative des domaines les uns par rapport aux autres. Une fois ce travail accompli, les chaînes latérales des acides aminés importants pour la cohésion des répétitions et la liaison à l'ADN devront être étudiées et si besoin est, correctement repositionnées. Ensuite viendra l'étude des chaînes latérales des autres résidus.

Mais si on considère que la structure support utilisée pour la modélisation appartient à un organisme, la souris, trop éloigné du parasite *P. falciparum*, on peut donc penser qu'elle est très proche mais pas identique à la structure réelle du domaine de liaison à l'ADN de

PfMyb1. Le manque de diversité en terme de structures de protéines Myb pose un réel problème. En effet, sur la vingtaine de structures présentes dans la base de données PDB (de une à trois répétitions), seules deux ne proviennent pas de la protéine c-Myb de souris. Il s'agit de la protéine B-Myb de poulet et de la protéine v-Myb du virus de la myéloblastose aviaire, cette dernière étant connue pour ne pas interagir avec la protéine C/EBPβ. Il faudrait donc considérer que les modèles générés sont des structures proches de la réalité mais qu'il est indispensable d'avoir des arguments biologiques pour savoir (i) quels sont exactement les acides aminés impliqués dans la liaison à l'ADN mais aussi quels sont les nucléotides impliqués dans cette liaison, (ii) quels sont les acides aminés importants pour maintenir la structure en Hélice-Tour-Hélice. Ces arguments biologiques permettront alors d'introduire des contraintes pour obtenir un meilleur modèle en combinant modélisation par homologie et modélisation *ab initio*.

DISCUSSION ET PERSPECTIVES

Chez les eucaryotes, l'expression des gènes en fonction d'un programme de développement passe par l'expression coordonnée de plusieurs familles de messagers, cette expression étant contrôlée tout au moins en partie au niveau de la transcription.

La transcription est régulée à plusieurs niveaux : des protéines reconnaissent des structures adoptées par la chromatine qu'elles remodèlent ce qui permet de décondenser la double hélice. Une fois la double hélice décondensée, des facteurs de transcription viennent se fixer de manière séquence-spécifique sur de petits éléments *cis*-régulateurs situés dans les promoteurs des gènes qui doivent être exprimés (voir Figure 19, p. 71). La mosaïque et l'agencement de ces éléments en modules de régulation déterminent la palette de facteurs pouvant se lier à ces séquences cibles et ainsi interagir avec la machinerie basale de transcription. Toutes ces interactions jouent un rôle crucial dans la régulation de la transcription conduisant à la modulation de l'expression des messagers et de leurs produits.

En effet, l'expression coordonnée de plusieurs ensembles de gènes, impliqués dans une étape du développement ou une fonction, va permettre aux cellules d'acquérir certaines caractéristiques morphogénétiques, de proliférer, de se différencier, de communiquer avec les cellules voisines, etc.

Lors de son développement complexe, *Plasmodium falciparum* subit de nombreuses modifications tant morphologiques que métaboliques (Figure 3, p. 23). La phase érythrocytaire est une étape cruciale pour le parasite. D'un côté, elle est responsable des symptômes de la maladie chez l'homme, car les fièvres sont provoquées par l'explosion des globules rouges dans lesquels le parasite a proliféré. De l'autre, elle est responsable de la dissémination du parasite *via* le moustique car la différenciation des mérozoïtes en gamétocytes, associée à un arrêt de la prolifération, est initiée dans les globules rouges.

Chaque étape de prolifération et de différenciation peut être caractérisée par l'expression préférentielle de certains gènes [39, 224] même s'il faut tout de même relativiser le lien entre la définition des stades d'un point de vue morphologique et du profil d'expression caractéristique des gènes. En effet, une expérience de transcriptome, dans laquelle l'expression relative des transcrits de *P. falciparum* a été mesurée heure par heure au cours du cycle érythrocytaire, a montré qu'il pourrait exister des vagues d'expression de transcrits qui ne sont pas restreintes aux stades définis morphologiquement [39]. Le contrôle fin de

l'expression transcriptionnelle des gènes, chez *P. falciparum*, apparaît alors comme une évidence. En effet, même si *P. falciparum* est un organisme encore assez mal connu d'un point de vue génétique (60% des phases ouvertes de lecture n'ont pas encore d'annotation), plusieurs études ont montré des caractéristiques communes aux autres eucaryotes comme l'organisation de la chromatine en nucléosomes [55], la structure des ARNm [220] ou encore la structure bipartite des promoteurs [70]. De plus, une fois le génome nucléaire du clone 3D7 entièrement séquencé, l'identification dans le génome de plus de 150 phases ouvertes de lecture (Annexes III.1 et III.2) codant des acteurs potentiels de la régulation transcriptionnelle des gènes, plaide pour un rôle crucial de ce contrôle dans le développement de *P. falciparum*. La diversité et la complexité des voies qui s'offrent à *P. falciparum* afin d'assurer la régulation transcriptionnelle de ses gènes va ainsi à l'encontre des conclusions sur une prétendue pauvreté en acteurs de la régulation transcriptionnelle [11, 68] tirées lors des premières analyses bioinformatiques de la séquence complète du génome de *P. falciparum*. L'analyse des structures secondaires des séquences de *P. falciparum* par HCA (Hydrophobic Cluster Analysis) pour la prédiction de nouvelles protéines orthologues semble une voie à suivre pour participer à l'annotation des protéines de *P. falciparum* car cette technique, couplée à PSI-Blast, a déjà permis à Isabelle Callebaut et ses collaborateurs d'identifier de nouveaux facteurs généraux de la transcription associés à l'ARN polymérase II [50]. Cependant, toutes ces phases ouvertes de lecture, qu'elles aient été annotées par nos soins ou qu'elles soient issues de la littérature, n'ont pas été étudiées moléculairement ; le rôle des protéines codées par ces phases ouvertes de lecture, dont la fonction est encore le plus souvent hypothétique, et leur implication dans le contrôle de l'initiation de la transcription restent donc encore à étudier.

Le groupe dans lequel je travaille tente d'appréhender les mécanismes impliqués dans le contrôle transcriptionnel de l'expression des gènes au cours du cycle érythrocytaire. Mon travail de thèse a donc été d'identifier et d'étudier les deux partenaires de la régulation transcriptionnelle chez *Plasmodium falciparum* dans le but de comprendre la première étape de la régulation des gènes lors de la phase érythrocytaire du développement du parasite.

Peu de facteurs de transcription, en dehors de la TBP et des ARN polymérases I, II et III qui sont des facteurs généraux de la transcription, ont été annotés et étudiés moléculairement

chez *P. falciparum*, à l'image du faible pourcentage (1,3%) de protéines assignées à la régulation de la transcription annoncé par Richard Coulson [68] contre 4% chez la levure *S. cerevisiae* qui possède pourtant un nombre de gènes équivalent. Cette différence n'implique pas nécessairement qu'un plus faible nombre de gènes parasitaires soit impliqué dans ces processus, mais souligne un domaine de la biologie du parasite où les connaissances sont encore limitées. La difficulté d'identifier des protéines impliquées dans la régulation de la transcription peut provenir du fait que *P. falciparum* possède de nombreuses insertions de faible complexité entre les domaines globulaires de ses protéines [11]. En effet, le nombre de gènes codant des protéines chez *P. falciparum* est équivalent à celui des levures *S. cerevisiae* et *S. pombe* alors que son génome est considérablement plus long (Tableau 3, p. 33). Cette différence de longueur se reflète en partie dans les protéines. La comparaison de protéines orthologues montre que les protéines de *P. falciparum* peuvent être jusqu'à 50% plus longues que les protéines de levure. L'analyse de cette différence de longueur montre une caractéristique remarquable des protéines du parasite : elles présentent beaucoup plus d'insertions prédites comme étant non globulaires que la plupart des eucaryotes. Ces insertions montrent une composition biaisée caractéristique car elles sont composées d'un ou de quelques acides aminés et correspondent très fréquemment à des homopolymères, le plus souvent composés d'asparagines (voir pour exemple les Figures 39 et 46). Cet enrichissement des protéines en régions riches en asparagine et de faible complexité fait que *Plasmodium* se distingue de la plupart des eucaryotes (à l'exception de *Dictyostelium discoideum*) pour lesquels les régions de faible complexité sont riches en glutamine et acide glutamique.

Nous avons recherché dans le génome de *P. falciparum* de nombreuses phases ouvertes de lecture codant des facteurs appartenant à toutes les superfamilles décrites dans TRANSFAC® (p. 62). Ceci nous a permis d'annoter plusieurs membres de deux principales familles de facteurs nucléaires et d'analyser la fonction biologique de certaines d'entre eux. Les premiers appartiennent à la famille HMGB et se lient à l'ADN sans spécificité de séquence mais avec une spécificité de structure ; ils participent au remodelage de la chromatine. Les deuxièmes, qui appartiennent à la famille Myb, sont spécifiques de séquences d'ADN et participent à la modulation, positive ou négative, du niveau de transcription de leurs gènes cibles.

PfHMGB1 & PfHMGB2, deux facteurs architecturaux.

Quatre phases ouvertes de lecture possédant un ou deux domaines 'HMG-box' ont été identifiées. Chez les eucaryotes et en particulier chez les vertébrés, les facteurs de type HMGB, de la classe des protéines à architecture β, ont été décrits pour leur rôle dans l'expression [48].

La plus longue d'entre elles, que nous avions appelée PfHMGB3 (PFL0290w) renferme, en plus de deux domaines 'HMG-box', un domaine similaire au domaine de liaison à l'ADN des protéines de type Myb. D'après L. Aravind et ses collaborateurs, il ne s'agirait donc pas d'une protéine HMGB mais du facteur général de transcription TFIIB B' [11].

PfHMGB1 et PfHMGB2 sont les protéines que nous avons le plus particulièrement étudiées parmi les quatre protéines annotées. Ces petites protéines d'environ 100 acides aminés sont essentiellement constituées de leur domaine de liaison à l'ADN de type 'HMG-box' précédé uniquement d'une région basique d'une vingtaine d'acides aminés. Une analyse phylogénétique des domaines 'HMG-box' (Figure 29, p. 121) ainsi que la présence de résidus caractéristiques (Figure 30, p. 124) nous ont permis de regrouper ces deux protéines plasmodiales avec les protéines architecturales HMGB de vertébrés, de plantes, de la drosophile mais surtout avec la protéine NHP6A de la levure *Saccharomyces cerevisiae* et la protéine NHP1 de *Babesia bovis*, un autre hématozoaire dont le génome est très riche en A+T.

Contrairement aux protéines HMGB des vertébrés qui possèdent deux domaines A et B en tandem, les protéines PfHMGB1 et PfHMGB2 ne possèdent qu'un seul domaine 'HMG-box' comme les protéines de plantes, de la drosophile ou de la levure. D'après une étude phylogénétique faite avec uniquement des boîtes A et des boîtes B, l'unique domaine 'HMG-box' des protéines plasmodiales semble plus proche de la boîte B que de la boîte A des protéines de vertébrés. Chez l'homme, il a été montré que la capacité de liaison à l'ADN cruciforme est associée à la boîte A [396], mais que la boîte B flanquée d'une région basique présente une activité de reconnaissance de l'ADN plus marquée [413].

Les deux facteurs ont été modélisés par homologie à la protéine NHP6A de *S. cerevisiae*. Le domaine de liaison à l'ADN des deux protéines a été facile à modéliser et est composé de trois hélices α repliées en forme de L qui se logent dans le petit sillon de l'ADN. En revanche, la partie N-terminale des deux protéines a posé plus de problèmes : cette région basique est censée entourer l'ADN pour venir se loger dans le grand sillon de la face opposée de l'hélice.

Comme elle ne contient pas de structures secondaires complexes et contraignantes, les résultats obtenus pour la modélisation de cette partie des deux protéines sont plus incertains.

Pour avoir des données qui permettraient de modéliser correctement cette partie N-terminale, il serait intéressant de faire appel à la biologie. Pour jouer leur rôle architectural, les facteurs HMGB doivent être capables de se lier à des structures d'ADN de manière spécifique et aussi d'induire une courbure de l'ADN facilitant l'assemblage de complexes nucléoprotéiques importants. Il a donc été montré que les deux protéines parasitaires avaient bien un rôle architectural car (i) elles sont capables d'interagir *in vitro* avec l'ADN cruciforme (Figure 2, p. 676 de l'Article 2) et (ii) d'induire la courbure d'un ADN linéaire (Figure 3, p. 677 de l'Article 2). Cependant, dans les deux cas, PFHMGB1 est plus efficace que PfHMGB2. Cette différence d'efficacité pourrait s'expliquer par la région N-terminale des deux protéines : alors que les deux protéines présentent 67% d'identité au niveau de leur domaine de liaison à l'ADN, elles ne présentent que 30% d'identité au niveau de leur partie N-terminale et ne sont pas de la même longueur. Une équipe américaine a mis en évidence deux blocs de résidus basiques situés dans la partie N-terminale de la protéine NHP6A qui sont importants pour l'efficacité d'interaction et de courbure de l'ADN ainsi que pour la stabilisation de la liaison ADN-protéine [412]. Ces deux blocs sont présents mais pas identiques dans les protéines plasmodiales (Figure 34, p. 130). Pour voir si ces deux blocs sont aussi indispensables à l'activité biologique des deux protéines parasitaires qu'ils le sont pour NHP6A, des expériences de mutagenèse et de délétion seraient indispensables. De plus, pour savoir si la région basique est responsable de la différence d'efficacité d'interaction et de courbure que l'on peut observer entre les deux protéines, il serait intéressant de créer des protéines hybrides composées de la partie N-terminale de PfHMGB1 et du domaine de liaison à l'ADN de PfHMGB2, et inversement, et de comparer leur efficacité d'interaction et de courbure avec les résultats déjà obtenus pour les protéines recombinantes.

Il est aujourd'hui raisonnable de penser que chez *P. falciparum*, les facteurs HMGB participent au remodelage de la chromatine grâce à leurs propriétés architecturales. Les protéines PfHMGB pourraient donc avoir un rôle essentiel dans le glissement du nucléosome sur l'ADN et par conséquent dans la modification de l'accessibilité des facteurs de transcription aux éléments de régulation comme chez les autres eucaryotes [380]. De cette manière, elles seraient capables de réguler la transcription. Des approches moléculaires telles

que celles effectuées pour la facteur PfMyb1 (inactivation des gènes *pfhmgb1* et *pfhmgb2* par interférence avec des ARN double brin et immunoprécipitation de la chromatine) devrait permettre de déterminer la fonction des facteurs PfHMGB dans le parasite et en particulier d'identifier leurs gènes cibles.

Les protéines PfHMGB1 et PfHMGB2 sont vraisemblablement impliquées dans la régulation des gènes de *P. falciparum*. Ceci a d'autant plus d'importance que les différences d'expression de ces facteurs en fonction du stade de développement suggèrent une intervention spécifique de PfHMGB1 dans la prolifération asexuée et de PfHMGB2 dans la différenciation sexuée (Fig. 4c, p. 677 de l'Article 2). En effet, la cinétique d'expression des protéines PfHMGB1 et PfHMGB2, similaire au profil d'expression des transcrits obtenu par le laboratoire d'Elizabeth Winzeler [224] montre que PfHMGB1 est abondante dans les cellules au stade asexué tandis que PfHMGB2 l'est dans les gamétocytes. La localisation de PfHMGB1 et PfHMGB2 dans le noyau des différents stades de développement confirme leur fonction nucléaire (Fig. 5, p. 679 de l'Article 2). Cependant, la présence de PfHMGB2 dans le cytoplasme des gamétocytes renforce l'hypothèse du rôle non redondant de ces protéines dans la cellule tout comme les différences observées vis-à-vis de leur interaction avec l'ADN. L'inactivation spécifique de l'un ou l'autre des deux gènes *pfhmgb1* et *pfhmgb2* pourrait alors nous informer sur la redondance de ces facteurs dans le parasite. Il est intéressant de noter que les protéines HMGB1 et HMGB2 humaines ont aussi une expression différente suivant le contexte cellulaire et le stade de développement et ne sont pas interchangeables en dépit de leur très grand pourcentage d'identité (80%) [328]. En effet, HMGB1 est une protéine ubiquitaire et les souris déficientes *hmgb1*$^{-/-}$ ne sont pas viables [51] tandis que HMGB2 est majoritairement exprimée dans les organes lymphoïdes et les testicules et les souris *hmgb2*$^{-/-}$ sont viables mais présentent une sévère baisse de la fertilité [328].

La fonction nucléaire des protéines HMGB est essentielle chez les eucaryotes mais la littérature porte un intérêt grandissant sur la double vie de ces protéines quant à leur rôle pro-inflammatoire [265]. En effet, les protéines HMGB libérées activement ou passivement dans le milieu extracellulaire sont de puissants activateurs de monocytes et déclenchent la production de cytokines comme le TNFα [390, 411] en faisant intervenir un autre domaine fonctionnel de la protéine appelé 'domaine stimulant le TNFα' [229]. Un domaine présentant 75 % de fortes similitudes avec ce domaine est présent dans les protéines plasmodiales. Ceci

suggère que les facteurs PfHMGB pourraient être aussi une cytokine cruciale impliquée dans la réponse inflammatoire à l'infection par *P. falciparum.*

La physiopathologie du neuropaludisme est encore mal comprise mais elle semble résulter de la séquestration des globules rouges infectés et des réponses pro-inflammatoires [105]. Des niveaux élevés de cytokines, en particulier de TNFα, ont été corrélés à la sévérité de la maladie [149, 216]. Plus récemment, une étude portant sur 16 enfants, dont 10 sont morts par neuropaludisme, a rapportée que le taux sérique de la protéine HMGB humaine pouvait être associé à la gravité de l'infection à *P. falciparum* [6]. Des expériences préliminaires montrent que les deux protéines PfHMGB1 et PfHMGB2 sont libérées par le parasite et seraient capables d'activer des monocytes primaires humains induisant la production de cytokines (Sylvie Briquet, en collaboration avec Vincent Maréchal). Ces premiers résultats sur l'activité cytokinique potentielle des facteurs PfHMGB ont d'autant plus d'importance que ces protéines pourraient jouer un rôle clé dans le développement de la pathologie cérébrale, le neuropaludisme. L'utilisation d'un modèle murin de neuropaludisme pourrait nous donner un élément de réponse quant à la fonction de ces facteurs *in vivo*. Il serait intéressant de déterminer si la neutralisation de la protéine HMGB du parasite murin *Plasmodium berghei* après immunisation ou injection d'anticorps pourrait entraver le développement de la pathologie cérébrale et influer ainsi sur la survie de souris infectée par *P. berghei*.

PfMyb1, un facteur de transcription spécifique de séquences d'ADN.

Trois phases ouvertes de lecture possédant deux ou trois domaines Myb ont été annotées. Les protéines Myb sont très conservées au cours de l'évolution et sont connues pour être impliquées dans la régulation de la croissance cellulaire et de la différenciation.

Avec le séquençage du génome de *P. falciparum*, une quarantaine de protéines contenant des domaines Myb a été annotée par le consortium, même si cette annotation ne se révèle pas toujours correcte. De plus, des protéines possédant un ou plusieurs domaines Myb peuvent ne pas être des facteurs de transcription Myb. En effet, la plus longue des trois protéines, que nous avions appelée PfMyb3 (PF10_0143) renferme, en plus de deux domaines Myb, un domaine en doigt de zinc de type ZZ et un domaine 'IMP dehydrogenase / GMP reductase'.

Qi Fan et ses collaborateurs ont étudié cette protéine [109] et il s'agit en fait d'un co-activateur transcriptionnel, PfADA2, qui s'associe *in vivo* avec l'HAT PfGCN5. Le complexe PfADA2-PfGCN5 serait impliqué dans le remodelage de la chromatine et donc dans la régulation de la transcription, à l'image de ce qui se passe chez la levure, l'homme et la drosophile.

Le facteur de transcription PfMyb1 possède un domaine de liaison à l'ADN composé de trois répétitions en tandem (R1, R2 et R3) situé en C-terminal de la protéine. Une modélisation par homologie au domaine de liaison à l'ADN de la protéine c-Myb de souris (Figure 57, p. 183) a montré que les domaines R1 et R2 sont composés de trois hélices qui se replient pour former une sorte de H (architecture en Hélice-Tour-Hélice). La répétition R3 possède une insertion de six acides aminés dans sa deuxième moitié et n'a donc pu être modélisée dans sa totalité. Les modèles obtenus par homologie ne présentent donc que deux hélices pour la répétition R3.

Cette protéine est capable de se lier à des éléments de régulation qui lui sont spécifiques et qui répondent au consensus des sites de fixation de protéines Myb d'autres eucaryotes (Figure 2b-e, p. 161 de l'Article 3). Les éléments de régulation de type *myb* ont d'ailleurs été recherchés dans les promoteurs de tous les gènes de *P. falciparum*. Il apparaît que ces éléments de régulation n'apparaissent pas par hasard dans les régions intergéniques quand les résultats sont comparés à ceux obtenus avec des séquences aléatoires de même composition que les promoteurs. Cependant, les éléments de régulation *myb* apparaissent aussi dans les séquences codantes. *In vivo*, les facteurs de transcription se fixent préférentiellement sur les éléments de régulation situés dans les régions intergéniques, même si les déterminants de cette sélectivité restent encore inconnus [234]. Il est donc possible que le facteur PfMyb1 possède aussi cette spécificité même s'il ne faut pas oublier qu'un facteur de transcription, quel qu'il soit, peut aussi se fixer sur un élément de régulation qui peut se situer à des milliers de pb du site d'initiation de la transcription et donc se trouver dans une phase ouverte de lecture.

D'après les expériences de retardement sur gel (Figure 2b-e, p. 161 de l'Article 3) où plusieurs bandes apparaissent (une majoritaire et deux minoritaires), il se pourrait que des partenaires interagissent avec la protéine PfMyb1 quand elle se lie à son site de fixation. La littérature rapporte l'existence de nombreux partenaires pour les protéines Myb. Parmi eux,

la protéine C/EBP qui appartient à la famille des protéines à domaine basique et plus particulièrement à la catégorie des protéines portant une agrafe à leucines (Figure 15a, p. 65). A ce jour, dans la base de données PlasmoDB, 44 protéines annotées par le consortium de séquençage contiennent le motif appelé 'bZIP transcription factor' (Pfam : PF0070) caractéristique, entre autres, des protéines C/EBP, dont 33 sont décrites comme « hypothetical protein ». Mais une recherche de motifs dans ces 44 séquences protéiques avec l'outil MotifScan indique que seules 15 de ces séquences contiennent un motif 'bZIP transcription factor' mais qui n'est pas statistiquement sûr et qui pourrait donc être un faux positif. MotifScan conseille d'ailleurs d'apporter des preuves biologiques pour déterminer s'il s'agit d'un vrai ou d'un faux positif. Au vu des résultats, quelque peu décevants, de la recherche dans le génome de séquences codant des protéines C/EBP qui peuvent interagir avec PfMyb1, il faut tout de même garder en mémoire que ce n'est pas parce qu'un motif précis n'est pas trouvé dans une protéine qu'il n'existe pas. J'en veux pour preuve les répétitions R1 et R3 de la protéine PfMyb1 qui n'ont pas été identifiées par le programme MotifScan mais qui semblent bel et bien exister et être fonctionnelles. C'est pourquoi l'analyse seule des séquences protéiques ne suffit plus et que des méthodes, telles que HCA, intégrant des informations structurales, peuvent apporter de nouvelles connaissances sur la biologie du parasite.

La coopération entre la protéine PfMyb1 et une protéine de la famille C/EBP peut nécessiter la présence d'un élément de régulation de type *c/ebp* : sa position par rapport au site de fixation de PfMyb1 sera alors déterminante pour l'interaction protéine-protéine, même si l'interaction entre les deux protéines peut se faire en l'absence d'ADN [366]. Des modules de régulation composés d'éléments de régulation de type *myb* et *c/ebp* situés à moins de 100 nucléotides l'un de l'autre ont alors été recherchés dans les promoteurs de tous les gènes de *P. falciparum*. Tout comme pour l'élément de régulation *myb* seul, l'association des deux éléments en modules de régulation ne semble pas apparaître par hasard dans les régions intergéniques et apparaissent aussi dans les séquences codantes.

La recherche de partenaires potentiels de PfMyb1 est une ouverture très intéressante. En effet, tandis que la cinétique d'expression de la protéine PfMyb1 est similaire dans les clones 3D7 et F12 [141], la cinétique d'interaction de la protéine avec les éléments de régulation *myb* est différente dans les deux clones (Figure 2e, p. 161 de l'Article 3). Le clone F12, dérivant

d'une culture continue du clone 3D7, a perdu la capacité de produire des gamétocytes [3]. L'absence, dans le clone F12, d'un ou plusieurs partenaires indispensables à l'interaction de PfMyb1 avec son élément de régulation pourrait expliquer la disparition de cette interaction dans les stades anneau et schizonte du clone F12. Cette différence entre les deux clones pourrait ainsi être une des causes du phénotype du clone F12. Dans ce cadre, la recherche de partenaires de PfMyb1 prend d'autant plus d'importance qu'ils pourraient être impliqués dans le processus de gamétocytogénèse. Ces partenaires pourraient expliquer l'abolition de l'expression de certains gènes, dont *pfg27* [3]. Des techniques d'immunoprécipitation couplées à la spectrométrie de masse sont actuellement employées au laboratoire pour tenter d'identifier des partenaires de PfMyb1. Par ailleurs, l'action de modifications post-traductionnelles ne peut être exclue, ces dernières pouvant réguler la liaison de PfMyb1 à l'ADN.

Lorsque l'expression du transcrit *pfmyb1*, et donc de la protéine PfMyb1, est partiellement inhibée par interférence par ARN double brin, les parasites meurent prématurément lors de la transition du stade trophozoïte (stade où l'expression du transcrit est censée être maximale) au stade schizonte (Figure 1, p. 31 de l'Article 4). Cette mort serait en partie la conséquence de la baisse de représentativité du transcrit et de la protéine PfMyb1 dans les cultures traitées par l'ARN double brin *pfmyb1*. Ces données, ajoutées à la conservation du profil d'expression de *pfmyb1* dans plusieurs clones de *P. falciparum* ainsi qu'à la conservation de la séquence de PfMyb1 chez *Plasmodium spp*, sont autant d'arguments en faveur d'un rôle crucial de PfMyb1 pour la survie lors du cycle érythrocytaire de *P. falciparum*.

Huit gènes ont vu leur expression altérée par la sous-expression de PfMyb1 (Tableau 1, p. 35 de l'Article 4). Des expériences d'immunoprécipitation de la chromatine indiquent que six d'entre eux semblent être directement régulés par PfMyb1. Cette donnée indique qu'*ex vivo* la protéine PfMyb1 joue bien le rôle d'un facteur de transcription et en se liant à l'ADN dans le noyau du parasite, participe à la régulation transcriptionnelle de certains gènes. Les deux autres promoteurs n'ont pas pu être mis en évidence par ces expériences. Soit l'interaction ADN-protéine est trop faible pour être détectée par ce type d'expérience, soit la régulation de ces deux gènes est indirectement assurée par PfMyb1, contrôlant lui-même un

autre facteur de transcription responsable du contrôle de ces deux gènes, mais qui ne ferait pas partie de facteurs de transcription présents sur la puce.

Des éléments de régulation *myb* sont présents dans les promoteurs de cinq des six gènes directement régulés par PfMyb1, le promoteur du gène *pgk* (PFI1105w) étant l'exception. Cette donnée impliquerait que PfMyb1 possède d'autres moyens de se lier à l'ADN que par le biais d'un élément de régulation *myb*. Cette caractéristique a déjà été décrite pour une protéine de la famille Myb, Bas1 de *S. cerevisiae* [174], mais aussi pour d'autres protéines de ce même organisme contenant des domaines de liaison à l'ADN de type Myb [189, 213, 287]. De plus, la liaison de PfMyb1 au promoteur serait dépendante de la présence d'autres facteurs [366]. Les connaissances apportées par des expériences d'identification des sites de liaison de la protéine de type Myb dans tout le génome de *S. cerevisiae* (immunoprécipitation de la chromatine) devraient être applicables à PfMyb1. Dans cet article, les auteurs rapportent que Rap1 se fixe quasi-exclusivement à des promoteurs contenant son site de fixation, mais peut se lier à des loci ne possédant pas ce site [234]. Ceci peut s'expliquer par le fait que le domaine de liaison à l'ADN de la protéine Rap1 est différent du domaine de liaison à l'ADN de la protéine c-Myb de souris ou de la protéine Bas1 de la levure car il est composé d'un domaine Myb avec trois hélices α et d'un homéodomaine avec quatre hélices α [213]. Bien que ces deux domaines adoptent une conformation en Hélice-Tour-Hélice comme la protéine c-Myb de souris [285], leur façon d'interagir avec l'ADN est différente de ce que l'on peut observer dans la plupart des protéines Myb ou des protéines à homéodomaines comme Engrailed. Le domaine de liaison à l'ADN de la protéine PfMyb1 diffère lui aussi du domaine de liaison à l'ADN de la protéine c-Myb de souris dans le sens où la répétition R3 possède une insertion de six acides aminés dans sa deuxième moitié. Cette insertion peut entraîner un réarrangement dans la structure globulaire de la répétition R3. Cette particularité peut ainsi apporter à la protéine parasitaire une flexibilité qui lui permettrait de se fixer sur des promoteurs ne possédant pas son site de liaison spécifique. Et la coopération avec d'autres facteurs de transcription pourrait « encourager » cette flexibilité [234, 261].

Pour finir, des éléments de régulation spécifiques de *P. falciparum* ont été recherchés dans les huit promoteurs des gènes dont l'expression est altérée par une baisse de la disponibilité du facteur de transcription PfMyb1. Quatre motifs ont été mis en évidence.

Cependant, ces expériences ont été menées avec une puce à ADN thématique ne représentant que 150 gènes de *P. falciparum*. Afin de connaître tous les gènes cibles, directs ou indirects, de PfMyb1, des expériences de transcriptome en utilisant des puces à ADN comprenant la totalité des phases ouvertes de lecture du parasite sont en train d'être menées avec la participation du consortium Plateforme Post-génomique Pasteur *Plasmodium* (P4). Enfin, parmi les deux motifs identifiés dans les promoteurs des gènes ayant le même profil d'expression que *pfmyb1* et les quatre motifs identifiées dans les promoteurs des gènes cibles de PfMyb1, certains sont peut-être des sites de liaison pour des facteurs de transcription spécifiques du parasite et donc non identifiables par bioinformatique. De plus, des expériences d'immunoprécipitation de la chromatine au niveau de tout le génome nous renseigneraient sur tous les sites de liaisons occupés par PfMyb1 et nous permettraient de mieux comprendre les déterminants de la liaison de PfMyb1 à l'ADN.

L'absence de vaccin efficace et les différents systèmes d'échappement performant développés par *Plasmodium falciparum* ainsi que la recrudescence des résistances aux médicaments et aux insecticides imposent d'acquérir de nouvelles connaissances sur la biologie du parasite pour proposer des nouvelles stratégies capables de contrôler son développement.

Des études -allers et retours entre analyses bioinformatiques et biologiques- ont démontré l'existence d'authentiques facteurs HMGB et Myb impliqués dans la régulation transcriptionnelle lors de la phase érythrocytaire du développement du parasite. Cela représente une première étape vers la compréhension de la régulation transcriptionnelle des événements clés du cycle érythrocytaire de *P. falciparum* et permet d'ouvrir une nouvelle voie pour la lutte contre le paludisme.

REFERENCES BIBLIOGRAPHIQUES

1. Abrahamsen M.S., Templeton T.J., Enomoto S., *et al.*, *Complete genome sequence of the apicomplexan, Cryptosporidium parvum.* Science, 2004. **304**(5669): p. 441-5.
2. Affolter M., Percival-Smith A., Muller M., *et al.*, *DNA binding properties of the purified Antennapedia homeodomain.* Proc Natl Acad Sci U S A, 1990. **87**(11): p. 4093-7.
3. Alano P., Roca L., Smith D., *et al.*, *Plasmodium falciparum: parasites defective in early stages of gametocytogenesis.* Exp Parasitol, 1995. **81**(2): p. 227-35.
4. Alano P., Silvestrini F. and Roca L., *Structure and polymorphism of the upstream region of the pfg27/25 gene, transcriptionally regulated in gametocytogenesis of Plasmodium falciparum.* Mol Biochem Parasitol, 1996. **79**(2): p. 207-17.
5. Allain F.H., Yen Y.M., Masse J.E., *et al.*, *Solution structure of the HMG protein NHP6A and its interaction with DNA reveals the structural determinants for non-sequence-specific binding.* Embo J, 1999. **18**(9): p. 2563-79.
6. Alleva L.M., Yang H., Tracey K.J., *et al.*, *High mobility group box 1 (HMGB1) protein: possible amplification signal in the pathogenesis of falciparum malaria.* Trans R Soc Trop Med Hyg, 2005. **99**(3): p. 171-4.
7. Altschul S.F., Gish W., Miller W., *et al.*, *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
8. Amsterdam A., Nissen R.M., Sun Z., *et al.*, *Identification of 315 genes essential for early zebrafish development.* Proc Natl Acad Sci U S A, 2004. **101**(35): p. 12792-7.
9. Apweiler R., Bairoch A., Wu C.H., *et al.*, *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32 Database issue**: p. D115-9.
10. Arakawa H., Nagase H., Hayashi N., *et al.*, *Molecular cloning, characterization, and chromosomal mapping of a novel human gene (GTF3A) that is highly homologous to Xenopus transcription factor IIIA.* Cytogenet Cell Genet, 1995. **70**(3-4): p. 235-8.
11. Aravind L., Iyer L.M., Wellems T.E., *et al.*, *Plasmodium biology: genomic gleanings.* Cell, 2003. **115**(7): p. 771-85.
12. Archambault J., Milne C.A., Schappert K.T., *et al.*, *The deduced sequence of the transcription factor TFIIIA from Saccharomyces cerevisiae reveals extensive divergence from Xenopus TFIIIA.* J Biol Chem, 1992. **267**(5): p. 3282-8.
13. Arnone M.I. and Davidson E.H., *The hardwiring of development: organization and function of genomic regulatory systems.* Development, 1997. **124**(10): p. 1851-64.
14. Ashburner M., Ball C.A., Blake J.A., *et al.*, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
15. Ausio J., Abbott D.W., Wang X., *et al.*, *Histone variants and histone modifications: a structural perspective.* Biochem Cell Biol, 2001. **79**(6): p. 693-708.
16. Bailey T.L. and Elkan C., *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*, in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, R. Altman, D. Brutlag, P. Karp, *et al.*, Editors. 1994, AAAI Press: Menlo Park, CA. p. 28-36.
17. Bairoch A., Apweiler R., Wu C.H., *et al.*, *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33 Database Issue**: p. D154-9.
18. Banerji J., Rusconi S. and Schaffner W., *Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences.* Cell, 1981. **27**(2 Pt 1): p. 299-308.
19. Bannister L. and Mitchell G., *The ins, outs and roundabouts of malaria.* Trends Parasitol, 2003. **19**(5): p. 209-13.
20. Barinaga M., *Dimers direct development.* Science, 1991. **251**(4998): p. 1176-7.
21. Basta S.S., Soekirman, Karyadi D., *et al.*, *Iron deficiency anemia and the productivity of adult males in Indonesia.* Am J Clin Nutr, 1979. **32**(4): p. 916-25.
22. Bateman A., Coin L., Durbin R., *et al.*, *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.

23. Beauchamps P., Tourvieille B., Cesbron-Delauw M.F., *et al.*, *The partial sequence of the Plasmodium falciparum histone H4 gene.* Res Microbiol, 1997. **148**(3): p. 201-3.
24. Bennett B.J., Thompson J. and Coppel R.L., *Identification of Plasmodium falciparum histone 2B and histone 3 genes.* Mol Biochem Parasitol, 1995. **70**(1-2): p. 231-3.
25. Bentley D., *The mRNA assembly line: transcription and processing machines in the same factory.* Curr Opin Cell Biol, 2002. **14**(3): p. 336-42.
26. Berg J.M., *Potential metal-binding domains in nucleic acid binding proteins.* Science, 1986. **232**(4749): p. 485-7.
27. Berg J.M., *Zinc finger domains: hypotheses and current knowledge.* Annu Rev Biophys Biophys Chem, 1990. **19**: p. 405-21.
28. Berman H.M., Westbrook J., Feng Z., *et al.*, *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
29. Bi W., Wu L., Coustry F., *et al.*, *DNA binding specificity of the CCAAT-binding factor CBF/NF-Y.* J Biol Chem, 1997. **272**(42): p. 26562-72.
30. Bianchi M.E., Beltrame M. and Paonessa G., *Specific recognition of cruciform DNA by nuclear protein HMG1.* Science, 1989. **243**(4894 Pt 1): p. 1056-9.
31. Biedenkapp H., Borgmeyer U., Sippel A.E., *et al.*, *Viral myb oncogene encodes a sequence-specific DNA-binding activity.* Nature, 1988. **335**(6193): p. 835-7.
32. Billings P.C., Davis R.J., Engelsberg B.N., *et al.*, *Characterization of high mobility group protein binding to cisplatin-damaged DNA.* Biochem Biophys Res Commun, 1992. **188**(3): p. 1286-94.
33. Blackwood E.M. and Kadonaga J.T., *Going the distance: a current view of enhancer action.* Science, 1998. **281**(5373): p. 61-3.
34. Boeckmann B., Bairoch A., Apweiler R., *et al.*, *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
35. Bonne-Andrea C., Harper F., Sobczak J., *et al.*, *Rat liver HMG1: a physiological nucleosome assembly factor.* Embo J, 1984. **3**(5): p. 1193-9.
36. Boschet C., *Etude de la régulation transcriptionnelle des gènes de Plasmodium falciparum*, Rapport de DEA, 20 p., Université Denis Diderot - Paris VII, Paris.
37. Bowie J.U., Luthy R. and Eisenberg D., *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.
38. Bowman S., Lawson D., Basham D., *et al.*, *The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum [see comments].* Nature, 1999. **400**(6744): p. 532-8.
39. Bozdech Z., Llinas M., Pulliam B.L., *et al.*, *The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.* PLoS Biol, 2003. **1**(1): p. E5.
40. Breman J.G., *The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden.* Am J Trop Med Hyg, 2001. **64**(1-2 Suppl): p. 1-11.
41. Briggs M.R., Kadonaga J.T., Bell S.P., *et al.*, *Purification and biochemical characterization of the promoter-specific transcription factor, Sp1.* Science, 1986. **234**(4772): p. 47-52.
42. Brown G.V., Culvenor J.G., Crewther P.E., *et al.*, *Localization of the ring-infected erythrocyte surface antigen (RESA) of Plasmodium falciparum in merozoites and ring-infected erythrocytes.* J Exp Med, 1985. **162**(2): p. 774-9.
43. Bucher P., *Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.* J Mol Biol, 1990. **212**(4): p. 563-78.
44. Buratowski S., *The basics of basal transcription by RNA polymerase II.* Cell, 1994. **77**(1): p. 1-3.
45. Buratowski S., Hahn S., Guarente L., *et al.*, *Five intermediate complexes in transcription initiation by RNA polymerase II.* Cell, 1989. **56**(4): p. 549-61.
46. Burk O., Mink S., Ringwald M., *et al.*, *Synergistic activation of the chicken mim-1 gene by v-myb and C/EBP transcription factors.* Embo J, 1993. **12**(5): p. 2027-38.
47. Burk O., Worpenberg S., Haenig B., *et al.*, *tom-1, a novel v-Myb target gene expressed in AMV- and E26-transformed myelomonocytic cells.* Embo J, 1997. **16**(6): p. 1371-80.

48. Bustin M., *Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins.* Mol Cell Biol, 1999. **19**(8): p. 5237-46.
49. Bustin M., *Revised nomenclature for high mobility group (HMG) chromosomal proteins.* Trends Biochem Sci, 2001. **26**(3): p. 152-3.
50. Callebaut I., Prat K., Meurice E., *et al.*, *Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes.* BMC Genomics, 2005. **6**: p. 100.
51. Calogero S., Grassi F., Aguzzi A., *et al.*, *The lack of chromosomal protein Hmg1 does not disrupt cell growth but causes lethal hypoglycaemia in newborn mice.* Nat Genet, 1999. **22**(3): p. 276-80.
52. Canutescu A.A., Shelenkov A.A. and Dunbrack R.L., Jr., *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Sci, 2003. **12**(9): p. 2001-14.
53. Carballo M., Puigdomenech P. and Palau J., *DNA and histone H1 interact with different domains of HMG 1 and 2 proteins.* Embo J, 1983. **2**(10): p. 1759-64.
54. Carey M., *The enhanceosome and transcriptional synergy.* Cell, 1998. **92**(1): p. 5-8.
55. Cary C., Lamont D., Dalton J.P., *et al.*, *Plasmodium falciparum chromatin: nucleosomal organisation and histone-like proteins.* Parasitol Res, 1994. **80**(3): p. 255-8.
56. Chalkley G.E. and Verrijzer C.P., *DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator.* Embo J, 1999. **18**(17): p. 4835-45.
57. Chao K.M., Pearson W.R. and Miller W., *Aligning two sequences within a specified diagonal band.* Comput Appl Biosci, 1992. **8**(5): p. 481-7.
58. Cheesman S., Horrocks P., Tosh K., *et al.*, *Intraerythrocytic expression of topoisomerase II from Plasmodium falciparum is developmentally regulated.* Mol Biochem Parasitol, 1998. **92**(1): p. 39-46.
59. Cho Y., Gorina S., Jeffrey P.D., *et al.*, *Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations.* Science, 1994. **265**(5170): p. 346-55.
60. Churchill M.E., Jones D.N., Glaser T., *et al.*, *HMG-D is an architecture-specific protein that preferentially binds to DNA containing the dinucleotide TG.* Embo J, 1995. **14**(6): p. 1264-75.
61. Clark K.L., Halay E.D., Lai E., *et al.*, *Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5.* Nature, 1993. **364**(6436): p. 412-20.
62. Cockwell K.Y. and Giles I.G., *Software tools for motif and pattern scanning: program descriptions including a universal sequence reading algorithm.* Comput Appl Biosci, 1989. **5**(3): p. 227-32.
63. Coin F. and Egly J.M., *Ten years of TFIIH.* Cold Spring Harb Symp Quant Biol, 1998. **63**: p. 105-10.
64. Combet C., Blanchet C., Geourjon C., *et al.*, *NPS@: network protein sequence analysis.* Trends Biochem Sci, 2000. **25**(3): p. 147-50.
65. Conaway R.C. and Conaway J.W., *General initiation factors for RNA polymerase II.* Annu Rev Biochem, 1993. **62**: p. 161-90.
66. Corpet F., *Multiple sequence alignment with hierarchical clustering.* Nucleic Acids Res, 1988. **16**(22): p. 10881-90.
67. Cosma M.P., *Ordered recruitment: gene-specific mechanism of transcription activation.* Mol Cell, 2002. **10**(2): p. 227-36.
68. Coulson R.M., Hall N. and Ouzounis C.A., *Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum.* Genome Res, 2004. **14**(8): p. 1548-54.
69. Cox J.M., Hayward M.M., Sanchez J.F., *et al.*, *Bidirectional binding of the TATA box binding protein to the TATA box.* Proc Natl Acad Sci U S A, 1997. **94**(25): p. 13475-80.
70. Crabb B.S. and Cowman A.F., *Characterization of promoters and stable transfection by homologous and nonhomologous recombination in Plasmodium falciparum.* Proc Natl Acad Sci U S A, 1996. **93**(14): p. 7289-94.
71. Cramer P., Larson C.J., Verdine G.L., *et al.*, *Structure of the human NF-kappaB p52 homodimer-DNA complex at 2.1 A resolution.* Embo J, 1997. **16**(23): p. 7078-90.
72. Creedon K.A., Kaslow D.C., Rathod P.K., *et al.*, *Identification of a Plasmodium falciparum histone 2A gene.* Mol Biochem Parasitol, 1992. **54**(1): p. 113-5.

73. Cuff J.A., Clamp M.E., Siddiqui A.S., *et al.*, *JPred: a consensus secondary structure prediction server.* Bioinformatics, 1998. **14**(10): p. 892-3.
74. Curran T. and Franza B.R., Jr., *Fos and Jun: the AP-1 connection.* Cell, 1988. **55**(3): p. 395-7.
75. Dalrymple B.P. and Peters J.M., *Characterization of a cDNA clone from the haemoparasite Babesia bovis encoding a protein containing an "HMG-Box".* Biochem Biophys Res Commun, 1992. **184**(1): p. 31-5.
76. Danis M., Legros F., Thellier M., *et al.*, *Données actuelles sur le paludisme en France métropolitaine.* Med Trop (Mars), 2002. **62**(3): p. 214-8.
77. Davey C.A., Sargent D.F., Luger K., *et al.*, *Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution.* J Mol Biol, 2002. **319**(5): p. 1097-113.
78. Davie J.R., *Covalent modifications of histones: expression from chromatin templates.* Curr Opin Genet Dev, 1998. **8**(2): p. 173-8.
79. Davie J.R. and Murphy L.C., *Inhibition of transcription selectively reduces the level of ubiquitinated histone H2B in chromatin.* Biochem Biophys Res Commun, 1994. **203**(1): p. 344-50.
80. Davison B.L., Egly J.M., Mulvihill E.R., *et al.*, *Formation of stable preinitiation complexes between eukaryotic class B transcription factors and promoter sequences.* Nature, 1983. **301**(5902): p. 680-6.
81. de la Cruz X., Lois S., Sanchez-Molina S., *et al.*, *Do protein motifs read the histone code?* Bioessays, 2005. **27**(2): p. 164-75.
82. de Villiers J. and Schaffner W., *A small segment of polyoma virus DNA enhances the expression of a cloned beta-globin gene over a distance of 1400 base pairs.* Nucleic Acids Res, 1981. **9**(23): p. 6251-64.
83. Dechering K.J., Kaan A.M., Eling W., *et al.*, unpublished.
84. Dechering K.J., Kaan A.M., Mbacham W., *et al.*, *Isolation and functional characterization of two distinct sexual-stage- specific promoters of the human malaria parasite Plasmodium falciparum.* Mol Cell Biol, 1999. **19**(2): p. 967-78.
85. Dechering K.J., Thompson J., Dodemont H.J., *et al.*, *Developmentally regulated expression of pfs16, a marker for sexual differentiation of the human malaria parasite Plasmodium falciparum.* Mol Biochem Parasitol, 1997. **89**(2): p. 235-44.
86. DeLano W.L., *The PyMOL molecular graphics system.* On the World Wide Web http://www.pymol.org, 2002.
87. Deleage G., Blanchet C. and Geourjon C., *Protein structure prediction. Implications for the biologist.* Biochimie, 1997. **79**(11): p. 681-6.
88. Delves C.J., Alano P., Ridley R.G., *et al.*, *Expression of alpha and beta tubulin genes during the asexual and sexual blood stages of Plasmodium falciparum.* Mol Biochem Parasitol, 1990. **43**(2): p. 271-8.
89. Dickerson R.E., *The DNA helix and how it is read.* Sci Am, 1983. **249**(6): p. 94-111.
90. Dini P.W. and Lipsick J.S., *Oncogenic truncation of the first repeat of c-Myb decreases DNA binding in vitro and in vivo.* Mol Cell Biol, 1993. **13**(12): p. 7334-48.
91. Doerig C., Horrocks P., Coyle J., *et al.*, *Pfcrk-1, a developmentally regulated cdc2-related protein kinase of Plasmodium falciparum.* Mol Biochem Parasitol, 1995. **70**(1-2): p. 167-74.
92. Doerig C.M., Parzy D., Langsley G., *et al.*, *A MAP kinase homologue from the human malaria parasite, Plasmodium falciparum.* Gene, 1996. **177**(1-2): p. 1-6.
93. Doolittle R.F., *The grand assault.* Nature, 2002. **419**(6906): p. 493-4.
94. Dorn A., Bollekens J., Staub A., *et al.*, *A multiplicity of CCAAT box-binding proteins.* Cell, 1987. **50**(6): p. 863-72.
95. Douguet D. and Labesse G., *Easier threading through web-based comparisons and cross-validations.* Bioinformatics, 2001. **17**(8): p. 752-3.
96. Dow L.K., Jones D.N., Wolfe S.A., *et al.*, *Structural studies of the high mobility group globular domain and basic tail of HMG-D bound to disulfide cross-linked DNA.* Biochemistry, 2000. **39**(32): p. 9725-36.

97. Dubendorff J.W., Whittaker L.J., Eltman J.T., *et al., Carboxy-terminal elements of c-Myb negatively regulate transcriptional activation in cis and in trans.* Genes Dev, 1992. **6**(12B): p. 2524-35.
98. Dynan W.S. and Tjian R., *Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins.* Nature, 1985. **316**(6031): p. 774-8.
99. Eberharter A. and Becker P.B., *Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics.* EMBO Rep, 2002. **3**(3): p. 224-9.
100. Eckstein-Ludwig U., Webb R.J., Van Goethem I.D., *et al., Artemisinins target the SERCA of Plasmodium falciparum.* Nature, 2003. **424**(6951): p. 957-61.
101. Eichinger L., Pachebat J.A., Glockner G., *et al., The genome of the social amoeba Dictyostelium discoideum.* Nature, 2005. **435**(7038): p. 43-57.
102. Eisenberg D., Luthy R. and Bowie J.U., *VERIFY3D: assessment of protein models with three-dimensional profiles.* Methods Enzymol, 1997. **277**: p. 396-404.
103. Ellis J., Ozaki L.S., Gwadz R.W., *et al., Cloning and expression in E. coli of the malarial sporozoite surface antigen gene from Plasmodium knowlesi.* Nature, 1983. **302**(5908): p. 536-8.
104. Emerson B.M., *Specificity of gene regulation.* Cell, 2002. **109**(3): p. 267-70.
105. Engwerda C., Belnoue E., Gruner A.C., *et al., Experimental models of cerebral malaria.* Curr Top Microbiol Immunol, 2005. **297**: p. 103-43.
106. Erondu N.E. and Donelson J.E., *Differential expression of two mRNAs from a single gene encoding an HMG1-like DNA binding protein of African trypanosomes.* Mol Biochem Parasitol, 1992. **51**(1): p. 111-8.
107. Etienne J., *Biochimie génétique. Biologie moléculaire*. Masson ed, ed. Abrégés. 1995
108. Etzold T. and Argos P., *SRS--an indexing and retrieval tool for flat file data libraries.* Comput Appl Biosci, 1993. **9**(1): p. 49-57.
109. Fan Q., An L. and Cui L., *PfADA2, a Plasmodium falciparum homologue of the transcriptional coactivator ADA2 and its in vivo association with the histone acetyltransferase PfGCN5.* Gene, 2004. **336**(2): p. 251-61.
110. Farid R.S., Bianchi M.E., Falciola L., *et al., Differential binding of HMG1, HMG2, and a single HMG box to cisplatin-damaged DNA.* Toxicol Appl Pharmacol, 1996. **141**(2): p. 532-9.
111. Favier D. and Gonda T.J., *Detection of proteins that bind to the leucine zipper motif of c-Myb.* Oncogene, 1994. **9**(1): p. 305-11.
112. Felsenstein J., *Confidence limits on phylogenies: an approach using the bootstrap.* Evolution, 1985. **39**: p. 783-791.
113. Felsenstein J., *PHYLIP -- Phylogeny Inference Package (Version 3.2).* Cladistics, 1989. **5**: p. 164-166.
114. Felsenstein J., *PHYLIP -- Phylogeny Inference Package (Version 3.63).* Cladistics, 1989. **5**: p. 164-166.
115. Fickett J.W. and Hatzigeorgiou A.G., *Eukaryotic promoter recognition.* Genome Res, 1997. **7**(9): p. 861-78.
116. Finney M., *The homeodomain of the transcription factor LF-B1 has a 21 amino acid loop between helix 2 and helix 3.* Cell, 1990. **60**(1): p. 5-6.
117. Fire A., Samuels M. and Sharp P.A., *Interactions between RNA polymerase II, factors, and template leading to accurate transcription.* J Biol Chem, 1984. **259**(4): p. 2509-16.
118. Fletcher C., *The Plasmodium falciparum genome project.* Parasitol Today, 1998. **14**(9): p. 342-4.
119. Flores O., Lu H., Killeen M., *et al., The small subunit of transcription factor IIF recruits RNA polymerase II into the preinitiation complex.* Proc Natl Acad Sci U S A, 1991. **88**(22): p. 9999-10003.
120. Flores O., Lu H. and Reinberg D., *Factors involved in specific transcription by mammalian RNA polymerase II. Identification and characterization of factor IIH.* J Biol Chem, 1992. **267**(4): p. 2786-93.
121. Fong I.C., Zarrin A.A., Wu G.E., *et al., Functional analysis of the human RAG 2 promoter.* Mol Immunol, 2000. **37**(7): p. 391-402.
122. Foulkes N.S., Borrelli E. and Sassone-Corsi P., *CREM gene: use of alternative DNA-binding domains generates multiple antagonists of cAMP-induced transcription.* Cell, 1991. **64**(4): p. 739-49.

123. Fox B.A., Li W.B., Tanaka M., *et al.*, *Molecular characterization of the largest subunit of Plasmodium falciparum RNA polymerase I.* Mol Biochem Parasitol, 1993. **61**(1): p. 37-48.
124. Fraenkel E., Rould M.A., Chambers K.A., *et al.*, *Engrailed homeodomain-DNA complex at 2.2 A resolution: a detailed view of the interface and comparison with other engrailed structures.* J Mol Biol, 1998. **284**(2): p. 351-61.
125. Frampton J., Gibson T.J., Ness S.A., *et al.*, *Proposed structure for the DNA-binding domain of the Myb oncoprotein based on model building and mutational analysis.* Protein Eng, 1991. **4**(8): p. 891-901.
126. Franklin S.G. and Zweidler A., *Non-allelic variants of histones 2a, 2b and 3 in mammals.* Nature, 1977. **266**(5599): p. 273-5.
127. Fromont-Racine M., Senger B., Saveanu C., *et al.*, *Ribosome assembly in eukaryotes.* Gene, 2003. **313**: p. 17-42.
128. Furukawa Y., Piwnica-Worms H., Ernst T.J., *et al.*, *cdc2 gene expression at the G1 to S transition in human T lymphocytes.* Science, 1990. **250**(4982): p. 805-8.
129. Gajiwala K.S. and Burley S.K., *Winged helix proteins.* Curr Opin Struct Biol, 2000. **10**(1): p. 110-6.
130. Ganss B. and Jheon A., *Zinc finger transcription factors in skeletal development.* Crit Rev Oral Biol Med, 2004. **15**(5): p. 282-97.
131. Gardiner-Garden M. and Frommer M., *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.
132. Gardner M.J., Hall N., Fung E., *et al.*, *Genome sequence of the human malaria parasite Plasmodium falciparum.* Nature, 2002. **419**(6906): p. 498-511.
133. Gardner M.J., Tettelin H., Carucci D.J., *et al.*, *Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum.* Science, 1998. **282**(5391): p. 1126-32.
134. Gaskins C.J. and Hanas J.S., *Sequence variation in transcription factor IIIA.* Nucleic Acids Res, 1990. **18**(8): p. 2117-23.
135. Gaskins C.J., Smith J.F., Ogilvie M.K., *et al.*, *Comparison of the sequence and structure of transcription factor IIIA from Bufo americanus and Rana pipiens.* Gene, 1992. **120**(2): p. 197-206.
136. Gautier R., Camproux A.C. and Tuffery P., *SCit: web tools for protein side chain conformation analysis.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W508-11.
137. Gehring W.J., Muller M., Affolter M., *et al.*, *The structure of the homeodomain and its functional implications.* Trends Genet, 1990. **6**(10): p. 323-9.
138. Geiger J.H., Hahn S., Lee S., *et al.*, *Crystal structure of the yeast TFIIA/TBP/DNA complex.* Science, 1996. **272**(5263): p. 830-6.
139. Gerondakis S. and Bishop J.M., *Structure of the protein encoded by the chicken proto-oncogene c-myb.* Mol Cell Biol, 1986. **6**(11): p. 3677-84.
140. Ginsberg A.M., King B.O. and Roeder R.G., *Xenopus 5S gene transcription factor, TFIIIA: characterization of a cDNA clone and measurement of RNA levels throughout development.* Cell, 1984. **39**(3 Pt 2): p. 479-89.
141. Gissot M., *Etude de la régulation transcriptionnelle des gènes lors du cycle érythrocytaire de Plasmodium falciparum*, 210 p., PhD thesis, Université Pierre et Marie Curie - Paris VI, Paris.
142. Glover J.N. and Harrison S.C., *Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA.* Nature, 1995. **373**(6511): p. 257-61.
143. Golay J., Capucci A., Arsura M., *et al.*, *Expression of c-myb and B-myb, but not A-myb, correlates with proliferation in human hematopoietic cells.* Blood, 1991. **77**(1): p. 149-58.
144. Gomez-Skarmeta J.L., Rodriguez I., Martinez C., *et al.*, *Cis-regulation of achaete and scute: shared enhancer-like elements drive their coexpression in proneural clusters of the imaginal discs.* Genes Dev, 1995. **9**(15): p. 1869-82.
145. Gonda T., Sheiness D. and Bishop J., *Transcripts from the cellular homologs of retroviral oncogenes: distribution among chicken tissues.* Mol Cell Biol, 1982. **2**(6): p. 617-624.

146. Gonda T.J., Gough N.M., Dunn A.R., *et al.*, *Nucleotide sequence of cDNA clones of the murine myb proto-oncogene.* Embo J, 1985. **4**(8): p. 2003-8.
147. Grasser F.A., LaMontagne K., Whittaker L., *et al.*, *A highly conserved cysteine in the v-Myb DNA-binding domain is essential for transformation and transcriptional trans-activation.* Oncogene, 1992. **7**(5): p. 1005-9.
148. Grasser K.D., Teo S.H., Lee K.B., *et al.*, *DNA-binding properties of the tandem HMG boxes of high-mobility-group protein 1 (HMG1).* Eur J Biochem, 1998. **253**(3): p. 787-95.
149. Grau G.E., Taylor T.E., Molyneux M.E., *et al.*, *Tumor necrosis factor and disease severity in children with falciparum malaria.* N Engl J Med, 1989. **320**(24): p. 1586-91.
150. Griffiths A.J.F., Miller J.H., Suzuki D.T., *et al.*, *An introduction to genetic analysis*. 5th ed. 1993, New York: W.H. Freeman
151. Grosschedl R., Giese K. and Pagel J., *HMG domain proteins: architectural elements in the assembly of nucleoprotein structures.* Trends in Genetics, 1994. **10**(3): p. 94-100.
152. Grunstein M., *Histone acetylation in chromatin structure and transcription.* Nature, 1997. **389**(6649): p. 349-52.
153. Guehmann S., Vorbrueggen G., Kalkbrenner F., *et al.*, *Reduction of a conserved Cys is essential for Myb DNA-binding.* Nucleic Acids Res, 1992. **20**(9): p. 2279-86.
154. Guex N. and Peitsch M.C., *SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.* Electrophoresis, 1997. **18**(15): p. 2714-23.
155. Guo K., Anjard C., Harwood A., *et al.*, *A myb-related protein required for culmination in Dictyostelium.* Development, 1999. **126**(12): p. 2813-22.
156. Hahn S., *Structure and mechanism of the RNA polymerase II transcription machinery.* Nat Struct Mol Biol, 2004. **11**(5): p. 394-403.
157. Hall R., Osland A., Hyde J.E., *et al.*, *Processing, polymorphism, and biological significance of P190, a major surface antigen of the erythrocytic forms of Plasmodium falciparum.* Mol Biochem Parasitol, 1984. **11**: p. 61-80.
158. Hamada H. and Bustin M., *Hierarchy of binding sites for chromosomal proteins HMG 1 and 2 in supercoiled deoxyribonucleic acid.* Biochemistry, 1985. **24**(6): p. 1428-33.
159. Hanas J.S., Hazuda D.J., Bogenhagen D.F., *et al.*, *Xenopus transcription factor A requires zinc for binding to the 5 S RNA gene.* J Biol Chem, 1983. **258**(23): p. 14120-5.
160. Harrison S.C., *A structural taxonomy of DNA-binding domains.* Nature, 1991. **353**(6346): p. 715-9.
161. Harrison S.C. and Aggarwal A.K., *DNA recognition by proteins with the helix-turn-helix motif.* Annu Rev Biochem, 1990. **59**: p. 933-69.
162. Hebbes T.R., Clayton A.L., Thorne A.W., *et al.*, *Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain.* Embo J, 1994. **13**(8): p. 1823-30.
163. Hebbes T.R., Thorne A.W. and Crane-Robinson C., *A direct link between core histone acetylation and transcriptionally active chromatin.* Embo J, 1988. **7**(5): p. 1395-402.
164. Hedberg K., Shaffer N., Davachi F., *et al.*, *Plasmodium falciparum-associated anemia in children at a large urban hospital in Zaire.* Am J Trop Med Hyg, 1993. **48**(3): p. 365-71.
165. Hirtzlin J., Farber P.M. and Franklin R.M., *Isolation of a novel Plasmodium falciparum gene encoding a protein homologous to the Tat-binding protein family.* Eur J Biochem, 1994. **226**(2): p. 673-80.
166. Hisatake K., Nishimura T., Maeda Y., *et al.*, *Cloning and structural analysis of cDNA and the gene for mouse transcription factor UBF.* Nucleic Acids Res, 1991. **19**(17): p. 4631-7.
167. Ho B.K. and Brasseur R., *The Ramachandran plots of glycine and pre-proline.* BMC Struct Biol, 2005. **5**: p. 14.
168. Hoch M., Schroder C., Seifert E., *et al.*, *cis-acting control elements for Krüppel expression in the Drosophila embryo.* Embo J, 1990. **9**(8): p. 2587-95.
169. Hoffman S.L., Bancroft W.H., Gottlieb M., *et al.*, *Funding for malaria genome sequencing.* Nature, 1997. **387**(6634): p. 647.

170. Horrocks P., Dechering K. and Lanzer M., *Control of gene expression in Plasmodium falciparum.* Mol Biochem Parasitol, 1998. **95**(2): p. 171-81.
171. Horrocks P., Jackson M., Cheesman S., *et al.*, *Stage specific expression of proliferating cell nuclear antigen and DNA polymerase delta from Plasmodium falciparum.* Mol Biochem Parasitol, 1996. **79**(2): p. 177-82.
172. Horrocks P. and Kilbey B.J., *Physical and functional mapping of the transcriptional start sites of Plasmodium falciparum proliferating cell nuclear antigen.* Mol Biochem Parasitol, 1996. **82**(2): p. 207-15.
173. Houbaviy H.B., Usheva A., Shenk T., *et al.*, *Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator.* Proc Natl Acad Sci U S A, 1996. **93**(24): p. 13577-82.
174. Hovring I., Bostad A., Ording E., *et al.*, *DNA-binding domain and recognition sequence of the yeast BAS1 protein, a divergent member of the Myb family of transcription factors.* J Biol Chem, 1994. **269**(26): p. 17663-9.
175. Hu J., Banerjee A. and Goss D.J., *Assembly of b/HLH/z proteins c-Myc, Max, and Mad1 with cognate DNA: importance of protein-protein and protein-DNA interactions.* Biochemistry, 2005. **44**(35): p. 11855-63.
176. Huang X. and Miller W., *A time-efficient, linear-space local similarity algorithm.* Adv Appl Math, 1991. **12**: p. 337-57.
177. Hughes J.D., Estep P.W., Tavazoie S., *et al.*, *Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.* J Mol Biol, 2000. **296**(5): p. 1205-14.
178. Hughey R. and Krogh A., *Hidden Markov models for sequence analysis: extension and analysis of the basic method.* Comput Appl Biosci, 1996. **12**(2): p. 95-107.
179. Hulo N., Sigrist C.J., Le Saux V., *et al.*, *Recent improvements to the PROSITE database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D134-7.
180. Introna M., Golay J., Frampton J., *et al.*, *Mutations in v-myb alter the differentiation of myelomonocytic cells transformed by the oncogene.* Cell, 1990. **63**(6): p. 1287-1297.
181. Jamin N., Gabrielsen O.S., Gilles N., *et al.*, *Secondary structure of the DNA-binding domain of the c-Myb oncoprotein in solution. A multidimensional double and triple heteronuclear NMR study.* Eur J Biochem, 1993. **216**(1): p. 147-54.
182. Jenuwein T. and Allis C.D., *Translating the histone code.* Science, 2001. **293**(5532): p. 1074-80.
183. Johnson L., Mollah S., Garcia B.A., *et al.*, *Mass spectrometry analysis of Arabidopsis histone H3 reveals distinct combinations of post-translational modifications.* Nucleic Acids Res, 2004. **32**(22): p. 6511-8.
184. Jones D.T., *GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.* J Mol Biol, 1999. **287**(4): p. 797-815.
185. Jones D.T., Taylor W.R. and Thornton J.M., *The rapid generation of mutation data matrices from protein sequences.* Comput Appl Biosci, 1992. **8**(3): p. 275-82.
186. Jones P.A., *The DNA methylation paradox.* Trends Genet, 1999. **15**(1): p. 34-7.
187. Jones P.L., Veenstra G.J., Wade P.A., *et al.*, *Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription.* Nat Genet, 1998. **19**(2): p. 187-91.
188. Joshi M.B., Lin D.T., Chiang P.H., *et al.*, *Molecular cloning and nuclear localization of a histone deacetylase homologue in Plasmodium falciparum.* Mol Biochem Parasitol, 1999. **99**(1): p. 11-9.
189. Ju Q.D., Morrow B.E. and Warner J.R., *REB1, a yeast DNA-binding protein with many targets, is essential for growth and bears some resemblance to the oncogene myb.* Mol Cell Biol, 1990. **10**(10): p. 5226-34.
190. Kadonaga J.T. and Grunstein M., *Chromosomes and expression mechanisms. Chromatin: the packaging is the message.* Curr Opin Genet Dev, 1999. **9**: p. 129-131.
191. Kanei-Ishii C., MacMillan E.M., Nomura T., *et al.*, *Transactivation and transformation by Myb are negatively regulated by a leucine-zipper structure.* Proc Natl Acad Sci U S A, 1992. **89**(7): p. 3088-92.

192. Kanei-Ishii C., Nomura T., Ogata K., *et al.*, *Structure and function of the proteins encoded by the myb gene family.* Curr Top Microbiol Immunol, 1996. **211**: p. 89-98.
193. Kaptein R., *Zinc fingers.* Curr Opin Struct Biol, 1991. **1**(1): p. 63-70.
194. Karlin S. and Altschul S.F., *Applications and statistics for multiple high-scoring segments in molecular sequences.* Proc Natl Acad Sci U S A, 1993. **90**(12): p. 5873-7.
195. Karlin S. and Altschul S.F., *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.* Proc Natl Acad Sci U S A, 1990. **87**(6): p. 2264-8.
196. Kasinsky H.E., Lewis J.D., Dacks J.B., *et al.*, *Origin of H1 linker histones.* Faseb J, 2001. **15**(1): p. 34-42.
197. Kassavetis G.A., Braun B.R., Nguyen L.H., *et al.*, *S. cerevisiae TFIIIB is the transcription initiation factor proper of RNA polymerase III, while TFIIIA and TFIIIC are assembly factors.* Cell, 1990. **60**(2): p. 235-45.
198. Katinka M.D., Duprat S., Cornillot E., *et al.*, *Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi.* Nature, 2001. **414**(6862): p. 450-3.
199. Katzen A.L., Kornberg T.B. and Bishop J.M., *Isolation of the proto-oncogene c-myb from D. melanogaster.* Cell, 1985. **41**(2): p. 449-56.
200. Kelley L.A., MacCallum R.M. and Sternberg M.J., *Enhanced genome annotation using structural profiles in the program 3D-PSSM.* J Mol Biol, 2000. **299**(2): p. 499-520.
201. Kerpola T.K. and Curran T., *Transcription factor interactions : basics on zippers.* Curr Opin Struct Biol, 1991. **1**(1): p. 71-79.
202. Kim C.G. and Sheffery M., *Physical characterization of the purified CCAAT transcription factor, alpha-CP1.* J Biol Chem, 1990. **265**(22): p. 13362-9.
203. Kim J.L., Nikolov D.B. and Burley S.K., *Co-crystal structure of TBP recognizing the minor groove of a TATA element.* Nature, 1993. **365**(6446): p. 520-7.
204. Kingston R.E. and Narlikar G.J., *ATP-dependent remodeling and acetylation as regulators of chromatin fluidity.* Genes Dev, 1999. **13**(18): p. 2339-52.
205. Kissinger J.C., Brunk B.P., Crabtree J., *et al.*, *The Plasmodium genome database.* Nature, 2002. **419**(6906): p. 490-2.
206. Klempnauer K.H., Gonda T.J. and Bishop J.M., *Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene.* Cell, 1982. **31**(2 Pt 1): p. 453-63.
207. Klug A. and Rhodes D., *Zinc fingers: a novel protein fold for nucleic acid recognition.* Cold Spring Harb Symp Quant Biol, 1987. **52**: p. 473-82.
208. Knudsen S., *Promoter2.0: for the recognition of PolII promoter sequences.* Bioinformatics, 1999. **15**(5): p. 356-61.
209. Knuppel R., Dietze P., Lehnberg W., *et al.*, *TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins.* J Comput Biol, 1994. **1**(3): p. 191-8.
210. Kohn W.D., Mant C.T. and Hodges R.S., *Alpha-helical protein assembly motifs.* J Biol Chem, 1997. **272**(5): p. 2583-6.
211. Kolodrubetz D. and Burgum A., *Duplicated NHP6 genes of Saccharomyces cerevisiae encode proteins homologous to bovine high mobility group protein 1.* J Biol Chem, 1990. **265**(6): p. 3234-9.
212. Kolodrubetz D., Haggren W. and Burgum A., *Amino-terminal sequence of a Saccharomyces cerevisiae nuclear protein, NHP6, shows significant identity to bovine HMG1.* FEBS Lett, 1988. **238**(1): p. 175-9.
213. Konig P., Giraldo R., Chapman L., *et al.*, *The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA.* Cell, 1996. **85**(1): p. 125-36.
214. Kornberg R.D. and Lorch Y., *Chromatin-modifying and -remodeling complexes.* Curr Opin Genet Dev, 1999. **9**(2): p. 148-51.
215. Koshland D. and Strunnikov A., *Mitotic chromosome condensation.* Annu Rev Cell Dev Biol, 1996. **12**: p. 305-33.

216. Kwiatkowski D., Hill A.V., Sambou I.*, et al.*, *TNF concentration in fatal cerebral, non-fatal cerebral, and uncomplicated Plasmodium falciparum malaria.* Lancet, 1990. **336**(8725): p. 1201-4.
217. Labesse G. and Mornon J., *Incremental threading optimization (TITO) to help alignment and modelling of remote homologues.* Bioinformatics, 1998. **14**(2): p. 206-11.
218. Landschulz W.H., Johnson P.F. and McKnight S.L., *The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins.* Science, 1988. **240**(4860): p. 1759-64.
219. Lanzer M., de Bruin D. and Ravetch J.V., *A sequence element associated with the Plasmodium falciparum KAHRP gene is the site of developmentally regulated protein-DNA interactions.* Nucleic Acids Res, 1992. **20**(12): p. 3051-6.
220. Lanzer M., de Bruin D. and Ravetch J.V., *Transcription mapping of a 100 kb locus of Plasmodium falciparum identifies an intergenic region in which transcription terminates and reinitiates.* Embo J, 1992. **11**(5): p. 1949-55.
221. Lanzer M., Wertheimer S.P., de Bruin D.*, et al.*, *Plasmodium: control of gene expression in malaria parasites.* Exp Parasitol, 1993. **77**(1): p. 121-8.
222. Laskowski R.A., Chistyakov V.V. and Thornton J.M., *PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.* Nucleic Acids Res, 2005. **33 Database Issue**: p. D266-8.
223. Lawrence C.E., Altschul S.F., Bogouski M.S.*, et al.*, *Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment.* Science, 1993. **262**(208-214).
224. Le Roch K.G., Zhou Y., Blair P.L.*, et al.*, *Discovery of gene function by expression profiling of the malaria parasite life cycle.* Science, 2003. **301**(5639): p. 1503-8.
225. Lee T.I. and Young R.A., *Transcription of eukaryotic protein-coding genes.* Annu Rev Genet, 2000. **34**: p. 77-137.
226. Legube G. and Trouche D., *Regulating histone acetyltransferases and deacetylases.* EMBO Rep, 2003. **4**(10): p. 944-7.
227. Lehming N., Thanos D., Brickman J.M.*, et al.*, *An HMG-like protein that can switch a transcriptional activator to a repressor.* Nature, 1994. **371**(6493): p. 175-9.
228. Lewis B.A. and Reinberg D., *The mediator coactivator complex: functional and physical roles in transcriptional regulation.* J Cell Sci, 2003. **116**(Pt 18): p. 3667-75.
229. Li J., Kokkola R., Tabibzadeh S.*, et al.*, *Structural basis for the proinflammatory cytokine activity of high mobility group box 1.* Mol Med, 2003. **9**(1-2): p. 37-45.
230. Li J., McConkey G.A., Rogers M.J.*, et al.*, *Plasmodium: the developmentally regulated ribosome.* Exp Parasitol, 1994. **78**(4): p. 437-41.
231. Li W.B., Bzik D.J., Gu H.M.*, et al.*, *An enlarged largest subunit of Plasmodium falciparum RNA polymerase II defines conserved and variable RNA polymerase domains.* Nucleic Acids Res, 1989. **17**(23): p. 9621-36.
232. Li W.B., Bzik D.J., Tanaka M.*, et al.*, *Characterization of the gene encoding the largest subunit of Plasmodium falciparum RNA polymerase III.* Mol Biochem Parasitol, 1991. **46**(2): p. 229-39.
233. Li X.Y., Mantovani R., Hooft van Huijsduijnen R.*, et al.*, *Evolutionary variation of the CCAAT-binding transcription factor NF-Y.* Nucleic Acids Res, 1992. **20**(5): p. 1087-91.
234. Lieb J.D., Liu X., Botstein D.*, et al.*, *Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association.* Nat Genet, 2001. **28**(4): p. 327-34.
235. Lipsick J.S., *One billion years of Myb.* Oncogene, 1996. **13**(2): p. 223-35.
236. Lnenicek-Allen M., Read C.M. and Crane-Robinson C., *The DNA bend angle and binding affinity of an HMG box increased by the presence of short terminal arms.* Nucleic Acids Res, 1996. **24**(6): p. 1047-51.
237. Lo W.S., Gamache E.R., Henry K.W.*, et al.*, *Histone H3 phosphorylation can promote TBP recruitment through distinct promoter-specific mechanisms.* Embo J, 2005. **24**(5): p. 997-1008.
238. Lodish H., Baltimore D., Berk A.*, et al.*, *Molecular cell biology*. 3rd ed. 1995, New York: W. H. Freeman and Company

239. Longhurst H.J. and Holder A.A., *The histones of Plasmodium falciparum: identification, purification and a possible role in the pathology of malaria.* Parasitology, 1997. **114 (Pt 5)**: p. 413-9.
240. Longhurst H.J. and Holder A.A., *The sequence of Plasmodium falciparum histone H3.* Mol Biochem Parasitol, 1995. **69**(1): p. 111-3.
241. Luger K. and Richmond T.J., *The histone tails of the nucleosome.* Curr Opin Genet Dev, 1998. **8**(2): p. 140-6.
242. Luscher B., Christenson E., Litchfield D.W., *et al., Myb DNA binding inhibited by phosphorylation at a site deleted during oncogenic activation.* Nature, 1990. **344**(6266): p. 517-22.
243. Luthy R., Bowie J.U. and Eisenberg D., *Assessment of protein models with three-dimensional profiles.* Nature, 1992. **356**(6364): p. 83-5.
244. Ma P.C., Rould M.A., Weintraub H., *et al., Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation.* Cell, 1994. **77**(3): p. 451-9.
245. Majello B., Kenyon L.C. and Dalla-Favera R., *Human c-myb protooncogene: nucleotide sequence of cDNA and organization of the genomic locus.* Proc Natl Acad Sci U S A, 1986. **83**(24): p. 9636-40.
246. Mal A. and Harter M.L., *MyoD is functionally linked to the silencing of a muscle-specific regulatory gene prior to skeletal myogenesis.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1735-9.
247. Maniatis T., Goodbourn S. and Fischer J.A., *Regulation of inducible and tissue-specific gene expression.* Science, 1987. **236**(4806): p. 1237-45.
248. Manley J.L., Fire A., Cano A., *et al., DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract.* Proc Natl Acad Sci U S A, 1980. **77**(7): p. 3855-9.
249. Marks P.A., Richon V.M. and Rifkind R.A., *Histone deacetylase inhibitors: inducers of differentiation or apoptosis of transformed cells.* J Natl Cancer Inst, 2000. **92**(15): p. 1210-6.
250. Martin A.C., *http://www.bioinf.org.uk/software/profit/.*
251. Martin C. and Paz-Ares J., *MYB transcription factors in plants.* Trends Genet, 1997. **13**(2): p. 67-73.
252. Marti-Renom M.A., Stuart A.C., Fiser A., *et al., Comparative protein structure modeling of genes and genomes.* Annu Rev Biophys Biomol Struct, 2000. **29**: p. 291-325.
253. Masquilier D. and Sassone-Corsi P., *Transcriptional cross-talk: nuclear factors CREM and CREB bind to AP-1 sites and inhibit activation by Jun.* J Biol Chem, 1992. **267**(31): p. 22460-6.
254. Mathieu O., Yukawa Y., Prieto J.L., *et al., Identification and characterization of transcription factor IIIA and ribosomal protein L5 from Arabidopsis thaliana.* Nucleic Acids Res, 2003. **31**(9): p. 2424-33.
255. Matys V., Fricke E., Geffers R., *et al., TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic Acids Res, 2003. **31**(1): p. 374-8.
256. McAndrew M.B., Read M., Sims P.F., *et al., Characterisation of the gene encoding an unusually divergent TATA- binding protein (TBP) from the extremely A+T-rich human malaria parasite Plasmodium falciparum.* Gene, 1993. **124**(2): p. 165-71.
257. McKnight S.L., *Functional relationships between transcriptional control signals of the thymidine kinase gene of herpes simplex virus.* Cell, 1982. **31**(2 Pt 1): p. 355-65.
258. McNabb D.S., Tseng K.A. and Guarente L., *The Saccharomyces cerevisiae Hap5p homolog from fission yeast reveals two conserved domains that are essential for assembly of heterotetrameric CCAAT-binding factor.* Mol Cell Biol, 1997. **17**(12): p. 7008-18.
259. McStay B., Hu C.H., Pikaard C.S., *et al., xUBF and Rib 1 are both required for formation of a stable polymerase I promoter complex in X. laevis.* Embo J, 1991. **10**(8): p. 2297-303.
260. Mizuguchi K., Deane C.M., Blundell T.L., *et al., HOMSTRAD: a database of protein structure alignments for homologous families.* Protein Sci, 1998. **7**(11): p. 2469-71.
261. Moretti P., Freeman K., Coodly L., *et al., Evidence that a complex of SIR proteins interacts with the silencer and telomere-binding protein RAP1.* Genes Dev, 1994. **8**(19): p. 2257-69.
262. Morgenstern B., *DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment.* Bioinformatics, 1999. **15**(3): p. 211-8.

263. Morris J.F., Rauscher F.J., 3rd, Davis B., *et al., The myeloid zinc finger gene, MZF-1, regulates the CD34 promoter in vitro.* Blood, 1995. **86**(10): p. 3640-7.
264. Morrow B.E., Ju Q. and Warner J.R., *A bipartite DNA-binding domain in yeast Reb1p.* Mol Cell Biol, 1993. **13**(2): p. 1173-82.
265. Muller S., Scaffidi P., Degryse B., *et al., New EMBO members' review: the double life of HMGB1 chromatin protein: architectural factor and extracellular signal.* Embo J, 2001. **20**(16): p. 4337-40.
266. Murphy F.V.t., Sweet R.M. and Churchill M.E., *The structure of a chromosomal high mobility group protein-DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition.* Embo J, 1999. **18**(23): p. 6610-8.
267. Murre C., McCaw P.S. and Baltimore D., *A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins.* Cell, 1989. **56**(5): p. 777-83.
268. Myrset A.H., Bostad A., Jamin N., *et al., DNA and redox state induced conformational changes in the DNA-binding domain of the Myb oncoprotein.* Embo J, 1993. **12**(12): p. 4625-33.
269. Nair S.K. and Burley S.K., *X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors.* Cell, 2003. **112**(2): p. 193-205.
270. Nakagoshi H., Nagase T., Kanei-Ishii C., *et al., Binding of the c-myb proto-oncogene product to the simian virus 40 enhancer stimulates transcription.* J Biol Chem, 1990. **265**(6): p. 3479-83.
271. Nakai K. and Horton P., *PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization.* Trends Biochem Sci, 1999. **24**(1): p. 34-6.
272. Nakajima N., Horikoshi M. and Roeder R.G., *Factors involved in specific transcription by mammalian RNA polymerase II: purification, genetic specificity, and TATA box-promoter interactions of TFIID.* Mol Cell Biol, 1988. **8**(10): p. 4028-40.
273. Nan X., Ng H.H., Johnson C.A., *et al., Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex.* Nature, 1998. **393**(6683): p. 386-9.
274. Narayan V.A., Kriwacki R.W. and Caradonna J.P., *Structures of zinc finger domains from transcription factor Sp1. Insights into sequence-specific protein-DNA recognition.* J Biol Chem, 1997. **272**(12): p. 7801-9.
275. Ner S.S., Churchill M.E., Searles M.A., *et al., dHMG-Z, a second HMG-1-related protein in Drosophila melanogaster.* Nucleic Acids Res, 1993. **21**(18): p. 4369-71.
276. Ness S.A., Kowenz-Leutz E., Casini T., *et al., Myb and NF-M: combinatorial activators of myeloid genes in heterologous cell types.* Genes Dev, 1993. **7**(5): p. 749-59.
277. Ness S.A., Marknell A. and Graf T., *The v-myb oncogene product binds to and activates the promyelocyte-specific mim-1 gene.* Cell, 1989. **59**(6): p. 1115-25.
278. Nightingale K., Dimitrov S., Reeves R., *et al., Evidence for a shared structural role for HMG1 and linker histones B4 and H1 in organizing chromatin.* Embo J, 1996. **15**(3): p. 548-61.
279. Nikolov D.B., Chen H., Halay E.D., *et al., Crystal structure of a human TATA box-binding protein/TATA element complex.* Proc Natl Acad Sci U S A, 1996. **93**(10): p. 4862-7.
280. Nikolov D.B., Chen H., Halay E.D., *et al., Crystal structure of a TFIIB-TBP-TATA-element ternary complex.* Nature, 1995. **377**(6545): p. 119-28.
281. Nomura N., Takahashi M., Matsui M., *et al., Isolation of human cDNA clones of myb-related genes, A-myb and B-myb.* Nucleic Acids Res, 1988. **16**(23): p. 11075-89.
282. Nomura T., Sakai N., Sarai A., *et al., Negative autoregulation of c-Myb activity by homodimer formation through the leucine zipper.* J Biol Chem, 1993. **268**(29): p. 21914-23.
283. Oelgeschlager T., Chiang C.M. and Roeder R.G., *Topology and reorganization of a human TFIID-promoter complex.* Nature, 1996. **382**(6593): p. 735-8.
284. Ogata K., Hojo H., Aimoto S., *et al., Solution structure of a DNA-binding unit of Myb: a helix-turn-helix- related motif with conserved tryptophans forming a hydrophobic core.* Proc Natl Acad Sci U S A, 1992. **89**(14): p. 6428-32.
285. Ogata K., Morikawa S., Nakamura H., *et al., Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices.* Cell, 1994. **79**(4): p. 639-48.

286. Ogilvie M.K. and Hanas J.S., *Molecular biology of vertebrate transcription factor IIIA: cloning and characterization of TFIIIA from channel catfish oocytes.* Gene, 1997. **203**(2): p. 103-12.
287. Ohi R., Feoktistova A., McCann S., *et al.*, *Myb-related Schizosaccharomyces pombe cdc5p is structurally and functionally conserved in eukaryotes.* Mol Cell Biol, 1998. **18**(7): p. 4097-108.
288. Ohndorf U.M., Rould M.A., He Q., *et al.*, *Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins.* Nature, 1999. **399**(6737): p. 708-12.
289. Onate S.A., Prendergast P., Wagner J.P., *et al.*, *The DNA-bending protein HMG-1 enhances progesterone receptor binding to its target DNA sequences.* Mol Cell Biol, 1994. **14**(5): p. 3376-91.
290. Orengo C.A., Michie A.D., Jones S., *et al.*, *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-108.
291. Orengo C.A., Pearl F.M., Bray J.E., *et al.*, *The CATH Database provides insights into protein structure/function relationships.* Nucleic Acids Res, 1999. **27**(1): p. 275-9.
292. Otsuka H. and Van Haastert P.J., *A novel Myb homolog initiates Dictyostelium development by induction of adenylyl cyclase expression.* Genes Dev, 1998. **12**(11): p. 1738-48.
293. Pabo C.O. and Sauer R.T., *Transcription factors: structural families and principles of DNA recognition.* Annu Rev Biochem, 1992. **61**: p. 1053-95.
294. Pagni M., Ioannidis V., Cerutti L., *et al.*, *MyHits: a new interactive resource for protein annotation and domain identification.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W332-5.
295. Parseghian M.H. and Hamkalo B.A., *A compendium of the histone H1 family of somatic subtypes: an elusive cast of characters and their characteristics.* Biochem Cell Biol, 2001. **79**(3): p. 289-304.
296. Parvin J.D. and Young R.A., *Regulatory targets in the RNA polymerase II holoenzyme.* Curr Opin Genet Dev, 1998. **8**(5): p. 565-70.
297. Paull T.T. and Johnson R.C., *DNA looping by Saccharomyces cerevisiae high mobility group proteins NHP6A/B. Consequences for nucleoprotein complex assembly and chromatin condensation.* J Biol Chem, 1995. **270**(15): p. 8744-54.
298. Payet D. and Travers A., *The acidic tail of the high mobility group protein HMG-D modulates the structural selectivity of DNA binding.* J Mol Biol, 1997. **266**(1): p. 66-75.
299. Paz-Ares J., Ghosal D., Wienand U., *et al.*, *The regulatory c1 locus of Zea mays encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators.* Embo J, 1987. **6**(12): p. 3553-8.
300. Pearl F.M., Lee D., Bray J.E., *et al.*, *Assigning genomic sequences to CATH.* Nucleic Acids Res, 2000. **28**(1): p. 277-82.
301. Perkins M., *Stage-dependent processing and localization of a Plasmodium falciparum protein of 130,000 molecular weight.* Exp Parasitol, 1988. **65**(1): p. 61-8.
302. Peters C.W., Sippel A.E., Vingron M., *et al.*, *Drosophila and vertebrate myb proteins share two conserved regions, one of which functions as a DNA-binding domain.* Embo J, 1987. **6**(10): p. 3085-90.
303. Peterson C.L. and Laniel M.A., *Histones and histone modifications.* Curr Biol, 2004. **14**(14): p. R546-51.
304. Pieler T. and Bellefroid E., *Perspectives on zinc finger protein function and evolution--an update.* Mol Biol Rep, 1994. **20**(1): p. 1-8.
305. Pil P.M. and Lippard S.J., *Specific binding of chromosomal protein HMG1 to DNA damaged by the anticancer drug cisplatin.* Science, 1992. **256**(5054): p. 234-7.
306. Pizzi E. and Frontali C., *Low-complexity regions in Plasmodium falciparum proteins.* Genome Res, 2001. **11**(2): p. 218-29.
307. Pologe L.G. and Ravetch J.V., *A chromosomal rearrangement in a P. falciparum histidine-rich protein gene is associated with the knobless phenotype.* Nature, 1986. **322**(6078): p. 474-7.
308. Prestridge D.S., *Predicting Pol II promoter sequences using transcription factor binding sites.* J Mol Biol, 1995. **249**(5): p. 923-32.
309. Proudfoot N.J., Furger A. and Dye M.J., *Integrating mRNA processing with transcription.* Cell, 2002. **108**(4): p. 501-12.

310. Ptashne M. and Gann A., *Genes and Signals*. 2002, New York: Cold Spring Harbor Laboratory Press
311. Pugh B.F. and Tjian R., *Transcription from a TATA-less promoter requires a multisubunit TFIID complex.* Genes Dev, 1991. **5**(11): p. 1935-45.
312. Purves W.K., Sadava D., Orians G.H., *et al.*, *Life: the science of biology*. 7th ed. 2004, New York: W. H. Freeman and Company
313. Ralston R. and Bishop J.M., *The protein products of the myc and myb oncogenes and adenovirus E1a are structurally related.* Nature, 1983. **306**(5945): p. 803-6.
314. Ravetch J.V., Kochan J. and Perkins M., *Isolation of the gene for a glycophorin-binding protein implicated in erythrocyte invasion by a malaria parasite.* Science, 1985. **227**(4694): p. 1593-7.
315. Rawlings D.J., Fujioka H., Fried M., *et al.*, *Alpha-tubulin II is a male-specific protein in Plasmodium falciparum.* Mol Biochem Parasitol, 1992. **56**(2): p. 239-50.
316. Ray-Gallet D., Gérard A., Polo S., *et al.*, *Variations sur le thème du "code histone".* Med Sci, 2005. **21**: p. 384-389.
317. Read C.M., Cary P.D., Crane-Robinson C., *et al.*, *Solution structure of a DNA-binding domain from HMG1.* Nucleic Acids Res, 1993. **21**(15): p. 3427-36.
318. Reeck G.R. and Teller D.C., *High mobility group proteins: purification, properties and amino acid sequence comparisons*, in *Progress in non histone protein research*, I. Bekhor and C.C. Liew, Editors. 1985, CRC Press Inc: Boca Raton, Fl, USA. p. 1-22.
319. Reese M.G., *Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome.* Comput Chem, 2001. **26**(1): p. 51-6.
320. Refour P., *Bio-puce à ADN thématique de Plasmodium falciparum : mise au point et validation d'une technique de détection de deux radio-isotopes après hybridation différentielle sur lame de verre*, 266 p., PhD thesis, Université Pierre et Marie Curie - Paris VI, Paris.
321. Reinberg D. and Roeder R.G., *Factors involved in specific transcription by mammalian RNA polymerase II. Purification and functional analysis of initiation factors IIB and IIE.* J Biol Chem, 1987. **262**(7): p. 3310-21.
322. Rice P., Longden I. and Bleasby A., *EMBOSS: the European Molecular Biology Open Software Suite.* Trends Genet, 2000. **16**(6): p. 276-7.
323. Ridley R.G., *Malaria: to kill a parasite.* Nature, 2003. **424**(6951): p. 887-9.
324. Robson K.J. and Jennings M.W., *The structure of the calmodulin gene of Plasmodium falciparum.* Mol Biochem Parasitol, 1991. **46**(1): p. 19-34.
325. Roeder R.G., *Eukaryotic nuclear RNA polymerases*, in *RNA polymerase*, R. Losick and M. Chamberlin, Editors. 1976, Cold Spring Harbor Laboratory: New York. p. 285-329.
326. Romier C., Cocchiarella F., Mantovani R., *et al.*, *The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y.* J Biol Chem, 2003. **278**(2): p. 1336-45.
327. Ronchi A., Bellorini M., Mongelli N., *et al.*, *CCAAT-box binding protein NF-Y (CBF, CP1) recognizes the minor groove and distorts DNA.* Nucleic Acids Res, 1995. **23**(22): p. 4565-72.
328. Ronfani L., Ferraguti M., Croci L., *et al.*, *Reduced fertility and spermatogenesis defects in mice lacking chromosomal protein Hmgb2.* Development, 2001. **128**(8): p. 1265-73.
329. Rosson D. and Reddy E.P., *Nucleotide sequence of chicken c-myb complementary DNA and implications for myb oncogene activation.* Nature, 1986. **319**(6054): p. 604-6.
330. Roth F.P., Hughes J.D., Estep P.W., *et al.*, *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.* Nat Biotechnol, 1998. **16**(10): p. 939-45.
331. Roussel M., Saule S., Lagrou C., *et al.*, *Three new types of viral oncogene of cellular origin specific for haematopoietic cell transformation.* Nature, 1979. **281**(5731): p. 452-5.
332. Saitou N. and Nei M., *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.

333. Scherf A., Hilbich C., Sieg K., *et al., The 11-1 gene of Plasmodium falciparum codes for distinct fast evolving repeats.* Embo J, 1988. **7**(4): p. 1129-37.
334. Schleif R., *DNA looping.* Annu Rev Biochem, 1992. **61**: p. 199-223.
335. Schleif R., *Gene regulation: why should DNA loop?* Nature, 1987. **327**(6121): p. 369-70.
336. Schnitzler G.R., Sif S. and Kingston R.E., *A model for chromatin remodeling by the SWI/SNF family.* Cold Spring Harb Symp Quant Biol, 1998. **63**: p. 535-43.
337. Schnitzler P., Hug M., Handermann M., *et al., Identification of genes encoding zinc finger proteins, non-histone chromosomal HMG protein homologue, and a putative GTP phosphohydrolase in the genome of Chilo iridescent virus.* Nucleic Acids Res, 1994. **22**(2): p. 158-66.
338. Scholz B.D., Gross R., Schultink W., *et al., Anaemia is associated with reduced productivity of women workers even in less-physically-strenuous tasks.* Br J Nutr, 1997. **77**(1): p. 47-57.
339. Schulman D.B. and Setzer D.R., *Identification and characterization of transcription factor IIIA from Schizosaccharomyces pombe.* Nucleic Acids Res, 2002. **30**(13): p. 2772-81.
340. Scott M.P., Tamkun J.W. and Hartzell G.W., 3rd, *The structure and function of the homeodomain.* Biochim Biophys Acta, 1989. **989**(1): p. 25-48.
341. Shiff C., Checkley W., Winch P., *et al., Changes in weight gain and anaemia attributable to malaria in Tanzanian children living under holoendemic conditions.* Trans R Soc Trop Med Hyg, 1996. **90**(3): p. 262-5.
342. Shindyalov I.N. and Bourne P.E., *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Eng, 1998. **11**(9): p. 739-47.
343. Singh J. and Dixon G.H., *High mobility group proteins 1 and 2 function as general class II transcription factors.* Biochemistry, 1990. **29**(26): p. 6295-302.
344. Sinha S., Maity S.N., Lu J., *et al., Recombinant rat CBF-C, the third subunit of CBF/NFY, allows formation of a protein-DNA complex with CBF-A and CBF-B and with yeast HAP2 and HAP3.* Proc Natl Acad Sci U S A, 1995. **92**(5): p. 1624-8.
345. Sippl M.J., *Recognition of errors in three-dimensional structures of proteins.* Proteins, 1993. **17**(4): p. 355-62.
346. Slansky J.E. and Farnham P.J., *Introduction to the E2F family: protein structure and gene regulation.* Curr Top Microbiol Immunol, 1996. **208**: p. 1-30.
347. Smale S.T. and Kadonaga J.T., *The RNA polymerase II core promoter.* Annu Rev Biochem, 2003. **72**: p. 449-79.
348. Sneath P.H.A. and Snokal R.R., *Numerical toxonomy*. 1973, San Fransisco: W. H. Freeman and Company
349. Snow R.W., Craig M., Deichmann U., *et al., Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population.* Bull World Health Organ, 1999. **77**(8): p. 624-40.
350. Solovyev V. and Salamov A., *The Gene-Finder computer tools for analysis of human and model organism genome sequences*, in *Proceedings of the fifth international conference on intelligent systems for molecular biology*, T. Gaasterland, P. Karp, K. Karplus, *et al.*, Editors. 1997, AAAI Press: Menlo Park, CA. p. 294-302.
351. Sopta M., Carthew R.W. and Greenblatt J., *Isolation of three proteins that bind to mammalian RNA polymerase II.* J Biol Chem, 1985. **260**(18): p. 10353-60.
352. Soullier S., Jay P., Poulat F., *et al., Diversification pattern of the HMG and SOX family members during evolution.* J Mol Evol, 1999. **48**(5): p. 517-27.
353. Soullier S., Jay P., Poulat F., *et al., Diversification pattern of the HMG and SOX family members during evolution.* J Mol Evol, 1999. **48**(5): p. 517-27.
354. Spencer V.A. and Davie J.R., *Role of covalent modifications of histones in regulating gene expression.* Gene, 1999. **240**(1): p. 1-12.
355. Stahl H.D., Kemp D.J., Crewther P.E., *et al., Sequence of a cDNA encoding a small polymorphic histidine- and alanine-rich protein from Plasmodium falciparum.* Nucleic Acids Res, 1985. **13**(21): p. 7837-46.

356. Stegmaier P., Kel A.E. and Wingender E., *Systematic DNA-Binding Domain Classification of Transcription Factors.* Genome Inform Ser Workshop Genome Inform, 2004. **15**(2): p. 276-86.
357. Stewart M.D., Li J. and Wong J., *Relationship between histone H3 lysine 9 methylation, transcription repression, and heterochromatin protein 1 recruitment.* Mol Cell Biol, 2005. **25**(7): p. 2525-38.
358. Stober-Grasser U., Brydolf B., Bin X., *et al.*, *The Myb DNA-binding domain is highly conserved in Dictyostelium discoideum.* Oncogene, 1992. **7**(3): p. 589-96.
359. Struhl K., *Yeast transcriptional regulatory mechanisms.* Annu Rev Genet, 1995. **29**: p. 651-74.
360. Su X.Z. and Wellems T.E., *Sequence, transcript characterization and polymorphisms of a Plasmodium falciparum gene belonging to the heat-shock protein (HSP) 90 family.* Gene, 1994. **151**(1-2): p. 225-30.
361. Sucgang R., Chen G., Liu W., *et al.*, *Sequence and structure of the extrachromosomal palindrome encoding the ribosomal RNA genes in Dictyostelium.* Nucleic Acids Res, 2003. **31**(9): p. 2361-8.
362. Sudarsanam P., Iyer V.R., Brown P.O., *et al.*, *Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 2000. **97**(7): p. 3364-9.
363. Suzuki M. and Brenner S.E., *Classification of multi-helical DNA-binding domains and application to predict the DBD structures of sigma factor, LysR, OmpR/PhoB, CENP-B, Rapl, and Xy1S/Ada/AraC.* FEBS Lett, 1995. **372**(2-3): p. 215-21.
364. Taddei A., Ray-Gallet D. and Almounzi G., *Assemblage et remodelage : le nucléosome sous influence.* Med Sci, 2000. **16**: p. 603-610.
365. Tahirov T.H., Sasaki M., Inoue-Bungo T., *et al.*, *Crystals of ternary protein-DNA complexes composed of DNA-binding domains of c-Myb or v-Myb, C/EBPalpha or C/EBPbeta and tom-1A promoter fragment.* Acta Crystallogr D Biol Crystallogr, 2001. **57**(Pt 11): p. 1655-8.
366. Tahirov T.H., Sato K., Ichikawa-Iwata E., *et al.*, *Mechanism of c-Myb-C/EBP beta cooperation from separated sites on a promoter.* Cell, 2002. **108**(1): p. 57-70.
367. Takagi Y., Komori H., Chang W.H., *et al.*, *Revised subunit structure of yeast transcription factor IIH (TFIIH) and reconciliation with human TFIIH.* J Biol Chem, 2003. **278**(45): p. 43897-900.
368. Tan S., Hunziker Y., Sargent D.F., *et al.*, *Crystal structure of a yeast TFIIA/TBP/DNA complex.* Nature, 1996. **381**(6578): p. 127-51.
369. Tan S. and Richmond T.J., *Eukaryotic transcription factors.* Curr Opin Struct Biol, 1998. **8**(1): p. 41-8.
370. Tanese N., Pugh B.F. and Tjian R., *Coactivators for a proline-rich activator purified from the multisubunit human TFIID complex.* Genes Dev, 1991. **5**(12A): p. 2212-24.
371. Tanikawa J., Yasukawa T., Enari M., *et al.*, *Recognition of specific DNA sequences by the c-myb protooncogene product: role of three repeat units in the DNA-binding domain.* Proc Natl Acad Sci U S A, 1993. **90**(20): p. 9320-4.
372. Teo S.H., Grasser K.D. and Thomas J.O., *Differences in the DNA-binding properties of the HMG-box domains of HMG1 and the sex-determining factor SRY.* Eur J Biochem, 1995. **230**(3): p. 943-50.
373. Theunissen O., Rudt F., Guddat U., *et al.*, *RNA and DNA binding zinc fingers in Xenopus TFIIIA.* Cell, 1992. **71**(4): p. 679-90.
374. Thomas J.O. and Travers A.A., *HMG1 and 2, and related 'architectural' DNA-binding proteins.* Trends Biochem Sci, 2001. **26**(3): p. 167-74.
375. Thompson J.D., Higgins D.G. and Gibson T.J., *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
376. Thompson J.D., Plewniak F., Thierry J., *et al.*, *DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches.* Nucleic Acids Res, 2000. **28**(15): p. 2919-26.
377. Tice-Baldwin K., Fink G.R. and Arndt K.T., *BAS1 has a Myb motif and activates HIS4 transcription only in combination with BAS2.* Science, 1989. **246**(4932): p. 931-5.
378. Tjian R., *Molecular machines that control genes.* Sci Am, 1995. **272**(2): p. 54-61.

379. Tjian R. and Maniatis T., *Transcriptional activation: a complex puzzle with few easy pieces.* Cell, 1994. **77**(1): p. 5-8.
380. Travers A.A., *Priming the nucleosome: a role for HMGB proteins?* EMBO Rep, 2003. **4**(2): p. 131-6.
381. Tupler R., Perini G. and Green M.R., *Expressing the human genome.* Nature, 2001. **409**(6822): p. 832-3.
382. Turner B.M., *Cellular memory and the histone code.* Cell, 2002. **111**(3): p. 285-91.
383. Turner B.M., *Histone acetylation and an epigenetic code.* Bioessays, 2000. **22**(9): p. 836-45.
384. Usheva A. and Shenk T., *TATA-binding protein-independent initiation: YY1, TFIIB, and RNA polymerase II direct basal transcription on supercoiled template DNA.* Cell, 1994. **76**(6): p. 1115-21.
385. Usheva A. and Shenk T., *YY1 transcriptional initiator: protein interactions and association with a DNA site containing unpaired strands.* Proc Natl Acad Sci U S A, 1996. **93**(24): p. 13571-6.
386. Venturelli D., Travali S. and Calabretta B., *Inhibition of T-cell proliferation by a MYB antisense oligomer is accompanied by selective down-regulation of DNA polymerase alpha expression.* Proc Natl Acad Sci U S A, 1990. **87**(15): p. 5963-7.
387. Voronova A. and Baltimore D., *Mutations that disrupt DNA binding and dimer formation in the E47 helix-loop-helix protein map to distinct domains.* Proc Natl Acad Sci U S A, 1990. **87**(12): p. 4722-6.
388. Wagner C.R., Hamana K. and Elgin S.C., *A high-mobility-group protein and its cDNAs from Drosophila melanogaster.* Mol Cell Biol, 1992. **12**(5): p. 1915-23.
389. Wang A.H., Quigley G.J., Kolpak F.J., *et al.*, *Molecular structure of a left-handed double helical DNA fragment at atomic resolution.* Nature, 1979. **282**(5740): p. 680-6.
390. Wang H., Bloom O., Zhang M., *et al.*, *HMG-1 as a late mediator of endotoxin lethality in mice.* Science, 1999. **285**(5425): p. 248-51.
391. Wang W., Carey M. and Gralla J.D., *Polymerase II promoter activation: closed complex formation and ATP-driven start site opening.* Science, 1992. **255**(5043): p. 450-3.
392. Warren A.J., *Eukaryotic transcription factors.* Curr Opin Struct Biol, 2002. **12**(1): p. 107-14.
393. Waters A.P., *The ribosomal RNA genes of Plasmodium.* Adv Parasitol, 1994. **34**: p. 33-79.
394. Waters A.P., Syin C. and McCutchan T.F., *Developmental regulation of stage-specific ribosome populations in Plasmodium.* Nature, 1989. **342**(6248): p. 438-40.
395. Watson J.D. and Crick F.H., *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.
396. Webb M. and Thomas J.O., *Structure-specific binding of the two tandem HMG boxes of HMG1 to four-way junction DNA is mediated by the A domain.* J Mol Biol, 1999. **294**(2): p. 373-87.
397. Weber J.L., *Molecular biology of malaria parasites.* Exp Parasitol, 1988. **66**(2): p. 143-70.
398. Weideman C.A., Netter R.C., Benjamin L.R., *et al.*, *Dynamic interplay of TFIIA, TBP and TATA DNA.* J Mol Biol, 1997. **271**(1): p. 61-75.
399. Weil P.A., Segall J., Harris B., *et al.*, *Faithful transcription of eukaryotic genes by RNA polymerase III in systems reconstituted with purified DNA templates.* J Biol Chem, 1979. **254**(13): p. 6163-73.
400. Wellems T.E. and Howard R.J., *Homologous genes encode two distinct histidine-rich proteins in a cloned isolate of Plasmodium falciparum.* Proc Natl Acad Sci U S A, 1986. **83**(16): p. 6065-9.
401. Werner T., *Models for prediction and recognition of eukaryotic promoters.* Mamm Genome, 1999. **10**(2): p. 168-75.
402. Wesseling J.G., Snijders P.J., van Someren P., *et al.*, *Stage-specific expression and genomic organization of the actin genes of the malaria parasite Plasmodium falciparum.* Mol Biochem Parasitol, 1989. **35**(2): p. 167-76.
403. White J.H. and Kilbey B.J., *DNA replication in the malaria parasite.* Parasitol Today, 1996. **12**(4): p. 151-5.
404. Wingender E., Chen X., Fricke E., *et al.*, *The TRANSFAC system on gene expression regulation.* Nucleic Acids Res, 2001. **29**(1): p. 281-3.
405. Wolberger C., *Combinatorial transcription factors.* Curr Opin Genet Dev, 1998. **8**(5): p. 552-9.

406. Wood V., Gwilliam R., Rajandream M.A., *et al., The genome sequence of Schizosaccharomyces pombe.* Nature, 2002. **415**(6874): p. 871-80.
407. Wood V., Rutherford K., Ivens K.M., *et al., A re-annotation of the Saccharomyces cerevisiae genome.* Comp. Funct. Genom., 2001. **2**: p. 143-154.
408. Workman J.L. and Kingston R.E., *Alteration of nucleosome structure as a mechanism of transcriptional regulation.* Annu Rev Biochem, 1998. **67**: p. 545-79.
409. Wu C.H., Yeh L.S., Huang H., *et al., The Protein Information Resource.* Nucleic Acids Res, 2003. **31**(1): p. 345-7.
410. Wu Y., Sifri C.D., Lei H.H., *et al., Transfection of Plasmodium falciparum within human red blood cells.* Proc Natl Acad Sci U S A, 1995. **92**(4): p. 973-7.
411. Yang H., Wang H., Czura C.J., *et al., The cytokine activity of HMGB1.* J Leukoc Biol, 2005. **78**(1): p. 1-8.
412. Yen Y.M., Wong B. and Johnson R.C., *Determinants of DNA binding and bending by the Saccharomyces cerevisiae high mobility group protein NHP6A that are important for its biological activities. Role of the unique N terminus and putative intercalating methionine.* J Biol Chem, 1998. **273**(8): p. 4424-35.
413. Yoshioka K., Saito K., Tanabe T., *et al., Differences in DNA recognition and conformational change activity between boxes A and B in HMG2 protein.* Biochemistry, 1999. **38**(2): p. 589-95.
414. Yudkovsky N., Ranish J.A. and Hahn S., *A transcription reinitiation intermediate that is stabilized by activator.* Nature, 2000. **408**(6809): p. 225-9.
415. Zappavigna V., Falciola L., Helmer-Citterich M., *et al., HMG1 interacts with HOX proteins and enhances their DNA binding and transcriptional activation.* Embo J, 1996. **15**(18): p. 4981-91.
416. Zawel L. and Reinberg D., *Initiation of transcription by RNA polymerase II: a multi-step process.* Prog Nucleic Acid Res Mol Biol, 1993. **44**: p. 67-108.
417. Zheng N., Fraenkel E., Pabo C.O., *et al., Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP.* Genes Dev, 1999. **13**(6): p. 666-74.
418. Zheng X.M., Moncollin V., Egly J.M., *et al., A general transcription factor forms a stable complex with RNA polymerase B (II).* Cell, 1987. **50**(3): p. 361-8.

ANNEXES

I - Phylogénie

I.1 - Abréviations du nom des organismes

At	*Arabidopsis thaliana*	arabette
Bb	*Babesia bovis*	parasite
Bt	*Bos taurus*	vache
Ce	*Caenorhabditis elegans*	nématode
Ch	*Capra hircus*	chèvre
CIV	*Chilo iridescent virus*	virus infectant les invertébrés
Cj	*Callithrix jacchus*	ouistiti
Cr	*Catharanthus roseus*	pervenche de Madagascar
Cs	*Chelydra serpentina*	tortue
Ct	*Chironomus tentans*	moucheron
Dm	*Drosophila melanogaster*	mouche
Dr	*Danio rerio*	poisson zèbre
Gg	*Gallus gallus*	poulet
Gm	*Glycine max*	soja
Gog	*Gorilla gorilla*	gorille
Hl	*Hylobates lar*	gibbon
Hs	*Homo sapiens*	homme
Hv	*Hordeum vulgare*	orge
In	*Ipomoea nil*	liseron
Mca	*Mus caroli*	souris
Mce	*Mus cervicolor*	souris
Mco	*Mus cookii*	souris
Mh	*Mastomys hildibrantii*	rongeur
Mma	*Mus macedonicus*	souris
Mm	*Mus musculus*	souris
Mp	*Mus pahari*	souris
Ms	*Mus spretus*	souris
Nc	*Neurospora crassa*	levure
Oa	*Ovis aries*	mouton
Om	*Oncorhynchus mykiss*	truite arc-en-ciel
Os	*Oryza sativa*	riz
Pa	*Podospora anserina*	levure
Pb	*Plasmodium berghei*	parasite
Pf	*Plasmodium falciparum*	parasite
Pk	*Plasmodium knowlesi*	parasite
Pp	*Pongo pygmaeus*	orang-outan
Ps	*Pisum sativum*	pois
Pt	*Pan troglodytes*	chimpanzé
Pv	*Plasmodium vivax*	parasite
Py	*Plasmodium yoelii*	parasite
Rn	*Rattus norvegicus*	rat
Sc	*Saccharomyces cerevisiae*	levure de boulanger
Sm	*Sminthopsis macroura*	marsupial
Sp	*Schizosaccharomyces pombe*	levure à fission
Ss	*Sus scrofa*	porc
Stp	*Strongylocentrotus purpuratus*	oursin
Tt	*Tetrahymena thermophila*	protozoaire cilié
Vf	*Vicia faba*	fève
Xl	*Xenopus laevis*	xénope
Zm	*Zea mays*	maïs

I.2 - Numéros d'accession des séquences utilisées pour la phylogénie des HMG

Les numéros d'accession sont ceux de la base de données GenBank. Les noms des facteurs sont codés de la manière suivante : « abréviation_organisme.nom-facteur ».

AJ006222, Hs.SOX20 ; D13491, At.HMG ; D14314, Gg.HMG1 ; D14315, Gg.DEF1 ; D14718, Hs.HMG1R ; D30765, Xl.HMG2 ; D40599, Os.S2676 ; D41834, Os.S4664 ; D61688, Om.SOXLZ ; D61689, Mm.SOXLZ ; J02895, Ss.HMG2 ; L06453, Stp.HMG1 ; L07107, Mm.HMGX ; L07335, Hs.SOX ; L08048, Hs.HMG1PS ; L08814, Rn.CIIDBP ; L08825, Dm.SSRP1 ; L12169, In.HMG ; L16900, Sc.IXR1 ; L22300, CIV.HMG ; L28094, Hv.HMG ; L29542, Mh.SRY ; L29543, Mp.SRY ; L29544, Ms.SRY ; L29547, Mma.SRY ; L29548, Mce.SRY ; L29549, Mco.SRY ; L29551, Mm.SRYG ; L29552, Mca.SRY ; L32954, Om.HMGT2 ; L35032, Mm.SOX18 ; M21683, Ss.HMG1 ; M54787, Nc.MATA1 ; M61725, Rn.UBF ; M62810, Hs.MTTF-1 ; M63424, Tt.HMGC ; M73753, Sc.MTHMG ; M77023, Dm.HMGD ; M80574, Gg.HMG2 ; M81360, Bb.NHP1 ; M83665, Hs.HMG2 ; M86737, Hs.SSRP1 ; M87306, Tt.MLH ; M93254, Ct.HMG1B ; S46279, Sm.SRY ; S50213, Mus.SSRP1 ; S69429, Sm.SOX3 ; S83308, Hs.SOX5 ; U09551, Rn.HBP1 ; U12467, Gg.SOX3 ; U12532, Gg.SOX2 ; U12533, Gg.SOX9 ; U12534, Gg.SOX11 ; U13695, Hs.PMS1 ; U13881, Dm.DSP1 ; U15569, Bt.SRY ; U21933, Xl.HMG1 ; U23752, Hs.SOX11 ; U32614, Mm.SOX6 ; U35612, Hs.SOX22 ; U66141, Mm.SOX21 ; U68056, Dm.SOX70D ; X02666, Om.HMGT ; X07642, Sp.MTMC ; X12597, Hs.HMG1 ; X12796, Bt.HMG1 ; X15317, Sc.NHP6A ; X15318, Sc.NHPB ; X53461, Hs.UBF ; X53772, Hs.SRY ; X55491, Mm.SRYA ; X56687, Hs.NOR90 ; X57561, Xl.UBF ; X58245, Gm.HMG ; X58282, Zm.HMG ; X58636, Mm.LEF-1 ; X59869, Hs.TCF1 ; X60458, Sc.ROX1 ; X60831, Mm.UBF ; X62870, Hs.TCF3 ; X62871, Hs.TCF4 ; X63463, Gg.HMG2A ; X64195, Pa.FPR1 ; X65657, Mm.SOX5 ; X65667, Dm.SOX14 ; X67668, Mm.HMG2 ; X70298, Mm.SOX4 ; X70683, Hs.SOX4 ; X71135, Hs.SOX3 ; X71139, Dm.HMGZ ; X76774, Ps.HMG ; X79821, Dr.SOX19 ; X81456, Dm.HMG ; X86380, Pt.SRY ; X86382, Gog.SRY ; X86383, Pp.SRY ; X86384, Hl.SRY ; X86386, Cj.SRY ; Y00463, Rn.HMG1 ; Z18958, Mm.SOX9 ; Z21703, Vf.HMG ; Z28410, Cr.HMG ; Z30265, Oa.SRY ; Z30646, Ch.SRY ; Z31299, Mm.TEST750 ; Z46629, Hs.SOX9 ; Z46727, Sc.YD9395.07.

I.3 - Alignement utilisé pour la phylogénie des facteurs HMG

Les noms des domaines 'HMG-box' correspondent aux noms des facteurs, suivis du numéro de domaine, pour les protéines ayant plusieurs domaines 'HMG-box'.

```
Hs.MTTF1_2     PKRPRSAYNV YVAERFQEAK GDSPQEKLKT VK------EN WKNLSDSEKE LYIQHAKEDE TRYHNEMKSW E
Mm.HMGX_2      PKRPRSAYNI YVSESFQEAK DDSAQGKLKL VN------EA WKNLSPEEKQ AYIQLAKDDR IRYDNEMKSW E
Hs.HMG1_1      PRGKMSSYAF FVQTCREEHK KKHPDASVNF SEFSKKCSER WKTMSAKEKG KFEDMAKADK ARYEREMKTY I
Ss.HMG1_1      PRGKMSSYAF FVQTCREEHK KKHPDASVNF SEFSKKCSER WKTMSAKEKG KFEDMAKADK ARYEREMKTY I
Bt.HMG1_1      PRGKMSSYAF FVQTCREEHK KKHPDASVNF SEFSKKCSER WKTMSAKEKG KFEDMAKADK ARYEREMKTY I
Rn.HMG1_1      PRGKMSSYAF FVQTCREEHK KKHPDASVNF SEFSKKCSER WKTMSAKEKG KFEDMAKADK ARYEREMKTY I
Hs.HMG1PS_1    PTGKMSSYAF FVQTCREEHK KKHPDASVNF SEFSKKCSER WKTMSAKEKG KFEDMAKADK ARYEREMKTY I
Xl.HMG1_1      PRGKMSSYAY FVQTCREEHK KKHPDASVNF AEFSKKCSER WKTMSK-EKT KFEDMAKADK VRYEREMKSY I
Xl.HMG2_1      PRGKMSSYAY FVQTCREEHK KKHPDTSVNF SDFSKKCSER WKSMSAKEKG KFEDLAKGDK ARYEREMKTY I
Om.HMGT_1      PRGKMSSYAY FVQTRREEHK KKHPEASVNF SEFSKKCSER WKTMSAKEKG KFEDLAKLDK VRYEREMRSY I
Hs.HMG2_1      PRGKMSSYAF FVQTCREEHK KKHPDSSVNF AEFSKKCSER WKTMSAKEKS KFEDMAKSDK ARYDREMKNY V
Ss.HMG2_1      PRGKMSSYAF FVQTCREEHK KKHPDSSVNF AEFSKKCSER WKTMSAKEKS KFEDMAKSDK ARYDREMKNY V
Gg.HMG2_1      PRGKMSSYAY FVQTCPREHK KKHPDSSVNF AEFSRKCSER WKTMSSKEKG KFEEMAKGDK ARYDREMKNY V
Mm.HMG2_1      PLGKMSSYAF FVQTCREEHK KKHPNSSVNF AEISKKCSKR WKTMSAKENS KFEDLAKSDK ACYYREMKNY V
Om.HMGT2_1     PKGKTSSYAF FVATCREEHK KKHPGTSVNF SEFSKKCSER WRTMSAKEKV KFEDMAKGDK VRYDKDMKGY V
Gg.HMG1_1      PKGKMSAYAF FVQTCREEHK K-NPEVPVNF AEFSKKCSER WKTMSSKEKA KFDEMAKADK VRYDREMKDY G
Gg.HMG2A_1     PKGKMSAYAF FVQTCREEHK KKNPEVPVNF AEFSKKCSER WKTMSSKEKA KFDEMAKADK VRYDREMKDY G
Mm.TEST_1      PKGKMSAYAF FVQTCREEHK KKNPEVPVNF AEFSKKCSER WKTMSSKEKS KFDEMAKADK VRYDREMKDY G
Dm.HMG_1       PRGRMTAYAY FVQTCREEHK KKHPDETVIF AEFSRKCAER WKTMVDKEKK RFHEMAEKDK QRYEAEMQNY V
Dm.DSP1_1      PRGRMTAYAY FVQTCREEHK KKHPDETVIF AEFSRKCAER WKTMVDKEKK RFHEMAEKDK QRYEAEMQNY V
Stp.HMG1_1     PRGRMSAYAY FVQDSRAEHG KNHPNSPVRF AEFSKDCSAR WKALEEKGKG VFHEKSMRDK VRYDREMQSY K
```

```
Hs.HMG1_2       PKRPPSAFFL FCSEYRPKIK GEHP--GLSI GDVAKKLGEM WNNTAADDKQ PYEKKAAKLK EKYEKDIAAY R
Bt.HMG1_2       PKRPPSAFFL FCSEYRPKIK GEHP--GLSI GDVAKKLGEM WNNTAADDKQ PYEKKAAKLK EKYEKDIAAY R
Rn.HMG1_2       PKRPPSAFFL FCSEYRPKIK GEHP--GLSI GDVAKKLGEM WNNTAADDKQ PYEKKAAKLK EKYEKDIAAY R
Hs.HMG1PS_2     PKRLPSAFFL FCSEYRPKIK GEHP--GLSI GDVAKKLGEM WNNTAADDKQ PYEKKAAKLK EKYEKDIAAY R
Ss.HMG1_2       PKRPPSAFFL FCSEYRPKIK GEHP--GLSI GDVAKKLGEM WNNTAADDKH PYEKKAAKLK EKYEKDIAAY R
Hs.HMG1R        PKRPPSAFFL FCSEYHPKIK GEQL--GLPI SDVVKKLGEM WNNTAAEDKQ PCEKKAAKLK EKYKKDIAAY -
Xl.HMG1_2       PKRPPSAFFL FCSDFRPKIK GEHP--GSTI GDIAKKLGEM WNNTATDDKL PYERRAAKLK EKYEKDVAAY R
Hs.HMG2_2       PKRPPSAFFL FCSEHRPKIK SEHP--GLSI GDTAKKLGEM WSEQSAKDKQ PYEQKAAKLK EKYEKDIAAY R
Ss.HMG2_2       PKRPPSAFFL FCSEHRPKIK SEHP--GLSI GDTAKKLGEM WSEQSAKDKQ PYEQKAAKLK EKYEKDIAAY R
Gg.HMG2_2       PKRPPSAFFL FCSEHRPKIK NDHP--GLSI GDTAKKLGEM WSEQSAKDKQ PYEQKAAKLK EKYEKDIAAY R
Mm.HMG2_2       PKRPPSAFCL FCSENRPKIK IEYP--GLSI GDTAKKLGEM WSEQSAKEKQ PYEQKAAKLK EKYEKDFAAY R
Xl.HMG2_2       PKRPPSAFFI FCSEHRPQIK SETP--GLSI GDTAKKLGEL WAEQTPKDKL PHEQKAAKLK EKYEKDVAAY R
Gg.HMG1_2       PKRPPSGFFL FCSEFRPKIK STNP--GISI GDVAKKLGEM WNNLSDGEKQ PYNNKAAKLK EKYEKDVADY K
Gg.HMG2A_2      PKRPPSAFFL FCSEFRPKIK STNP--GISI GDVAKKLGEM WNNLSDGEKQ PYNNKAAKLK EKYEKDVADY K
Mm.TEST_2       PKRPPSGFFL FCSEFRPKIK STNP--GISI GDVAKKLGEM WNNLSDNEKQ PYVTKAAKLK EKYEKDVADY K
Om.HMGT2_2      PKRPPSAFFV FCAEHRGRIK ADNP--GMGI GDIAKQLGLL WGKQTPKDKQ PHEAKAAKLK EKYEKDVAAY K
Om.HMGT_2       PKRPSSAFFI FCADFRPQVK GETP--GLSI GDVAKKLGEK WNNLTAEDKV PYEKKASRLK EKYEKDITAY R
Dm.HMG_2        PKRSLSAFFW FCNDERNKVK ALNP--EFGV GDIAKELGRK WSDVDPEVKQ KYESMAERDK ARYEREMTEY K
Dm.DSP1_2       PKRSLSAFFW FCNDERNKVK ALNP--EFGV GDIAKELGRK WSDVDPEVKQ KYESMAERDK ARYEREMTEY K
Stp.HMG1_2      PKRNLSAFFI FSGENRAAIK SVHP--NWSV GDIAKELAVR WRAMTAGEKI PFDKGAAKDK ERYIKAMAEY K
Zm.HMG          PKRAPSAFFV FMEEFRKEFK EKNPK-NKSV AAVGKAAGDR WKSLSESDKA PYVAKANKLK LEYNKAIAAY N
Os.S4664        PKRAPSAFFV FMEEFRKEFK EKNPK-NKSV AAVGKAAGDR WKSLTEADKA PYVAKANKLK AEYNKAIAAY N
Hv.HMG          PKRAPSAFFV FMGEFREEFK QKNPK-NKSV AAVGKAAGER WKSLSESEKA PFVAKANKLK GEYNKAIASY N
Ps.HMG          PKRPPSAFFV FMEDFRKQFK KGNAD-NKAV SAVGKAAGAK WKSMTEAEKA PYAAKAEKRK AEYEKSMKSY N
Gm.HMG          PKRPPSAFFV FMEEFRKVFN KEHPE-NKAV SAVGKAAGAK WKTMSDAEKA PYVAKSEKRK VEYEKNMRAY N
In.HMG          PKRPPSAFFV FMEDFRKTYK EKHPN-NKSV AVVGKAGGDK WKQLTAAEKA PFISKAEKRK QEYEKNLQAY N
Os.S2676        PKRPPSAFFV FMEQFRKDYK EKHPN-VKQV SVIGKAGGDK WKSMTDADKA PFVTKAEKLK XEYTKKIDAT T
Vf.HMG          PKRPPSAFFV FMADFREQYK KDHPN-NKSV AAVGKACGEE WKSLSEEEKA PYVDRALKKK EEYEITLQAY -
Hs.SSRP1        PKRPMSAYML WLNASREKIK SDHP--GISI TDLSKKAGEI WKGMSKEKKE EWDRKAEDAR RDYEKAMKEY E
Mus.SSRP1       PKRPMSAYML WLNASREKIK SDHP--GISI TDLSKKAGEI WKGMSKEKKE EWDRKAEDAR REYEKAMKEY E
Rn.CIIDBP       PKRPMSAYML WLNASREKIK SDHP--GISI TDLSKKAGEI WKGMSKEKKE EWDRKGEDAR REYEEAMKEY E
Gg.DEF1         PKRPMSAYML WLNANREKIK SDHP--GISI TDLSKKAGEL WKAMSKEKKE EWDRKAEDAK RDYEKAMKEY S
Dm.HMGD         PKRPLSAYML WLNSARESIK RENP--GIKV TEVAKRGGEL WRAMKD--KS EWEAKAAKAK DDYDRAVKEF E
Ct.HMG1B        PKRPLSAYML WLNSARESIK RENP--DFKV TEVAKKGGEL WRGLKD--KS EWEAKAATAK QNYIRALQEY E
Dm.HMGZ         PKRPLSAYML WLNETREQIK KDNP--GSKV TDIAKRGGEL WRGLKD--KT EWEQKAIKMK EEYNKAVKEY E
Dm.SSRP1        PKRATTAFML WLNDTRESYK RENP--GIKV TEIAKKGGEM WKELKD--KS KWEDAAAKDK QRYHDEMRNY K
Tt.HMGC         PKRPLSAFFL FKQHNYEQVK KENP--NAKI TELTSMIAEK WKAVGEKEKK KYETLQSEAK AKYEKALQAY E
At.HMG          PKRAMSGFMF FSQMERDNIK KEHPGIA--F GEVGKVLGDK CVKCLLMIKS HMKPRLQVDK QRYKDEISDY K
Cr.HMG          APKSISAFMF FSQTERENVK KDNPGIA--F TDVGKVLGDR CNKCQLRKKH LLKQRLVADK KRYTDEISNY K
Sc.NHP6A        PKRALSAYMF FANENRDIVR SENPDIT--F GQVGKKLGEK WKALTPEEKQ PYEAKAQADK KRYESEKELY N
Sc.NHPB         PKRGLSAYMF FANENRDIVR SENPDVT--F GQVGRILGER WKALTAEEKQ PYESKAQADK KRYESEKELY N
Pv.HMGB1        PKRSLSAYMF FAKEKRAEII SRDPDLSKDV ATVGKMIGEA WNKLDEREKA PYEKKAQEDK LRYEREKVEY A
Pk.HMGB1        PKRSLSAYMF FAKEKRAEII SRDPDLSKDV ATVGKMIGEA WNKLDEREKA PYEKKAQEDK VRYEREKVEY A
Py.HMGB1        PKRSLSAYMF FAKEKRAEII TRDPSLSKDV ATVGKMIGEA WNKLDEREKA PYEKKAQEDK IRYEKEKMEY A
Pb.HMGB1        PKRSLSAYMF FAKEKRAEII TRDPSLSKDV ATVGKMIGEA WNKLDEREKA PYEKKAQEDK IRYEKEKMEY A
Pf.HMGB1        PKRSLSAYMF FAKEKRAEII SKQPELSKDV ATVGKMIGEA WNKLGEKEKA PFEKKAQEDK LRYEKEKAEY A
Bb.NHP1         PKRALSSYMF FAKEKRVEII AENPEIAKDV AAIGKMIGAA WNALSDEEKK PYERMSDEDR VRYEREKAEY A
Py.HMGB2        PKRALSAYMF YVKDKRLEII QERPELAKEV AQVGKLIGEA WGQLTPAQKA PYEKKAELDK VRYSKEIEEY R
Pb.HMGB2        PKRALSAYMF YVKDKRLEII QERPELAKEV AQVGKLIGEA WGQLTPAQKA PYEKKAELDK VRYSKEIEEY R
Pf.HMGB2        PKRALSAYMF YVKDKRLEII KEKPELAKDV AQVGKLIGEA WGQLSPAQKA PYEKKAQLDK VRYSKEIEEY R
Pv.HMGB2        PKRALSAYMF YVKDKRLELI KERPELAKDV AQVGKLVGEA WGQLSAAQKT PYEKKAQLDK VRYSKEIEEY R
Pk.HMGB2        PKRALSAYMF YVKDKRLELI KEKPELARNV AQVGKLVGEA WGKLSAAQKT PYEKKAQLDK VRYSKEIEEY R
Hs.UBF_1        PKKPLTPYFR FFMEKRAKYA KLHP--EMSN LDLTKILSKK YKELPEKKKM KYIQDFQREK QEFERNLARF R
Mm.UBF_1        PKKPLTPYFR FFMEKRAKYA KLHP--EMSN LDLTKILSKK YKELPEKKKM KYIQDFQREK QEFERNLARF R
Rn.UBF_1        PKKPLTPYFR FFMEKRAKYA KLHP--EMSN LDLTKILSKK YKELPEKKKM KYIQDFQREK QEFERNLARF R
Hs.NOR90_1      PKKPLTPYFR FFMEKRAKYA KLHP--EMSN LDLTKILSKK YKELPEKKKM KYIQDFQREK QEFERNLARF R
Xl.UBF_1        PKKPLTPYFR FFMEKRAKYA KLHP--EMSN LDLTKILSKK YKELPEKKKM KYIQDFQREK QDFERNMAKF R
Hs.MTTF1_1      PKKPVSSYLR FSKEQLPIFK AQNP--DAKT TELIRRIAQR WRELPDSKKK IYQDAYRAEW QVYKEEISRF K
Mm.HMGX_1       PKKPMSSYLR FSTEQLPKFK AKHP--DAKL SELVRKIAAL WRELPEAEKK VYEADFKAEW KAYKEAVSKY K
Hs.UBF_4        PKRPVSAMFI FSEEKRRQLQ EERP--ELSE SELTRLLARM WNDLSEKKKA KYKAREAALK AQSERKPGGE R
Rn.UBF_4        PKRPVSAMFI FSEEKRRQLQ EERP--ELSE SELTRLLARM WNDLSEKKKA KYKAREAALK AQSERKPGGE R
Hs.NOR90_3      PKRPVSAMFI FSEEKRRQLQ EERP--ELSE SELTRLLARM WNDLSEKKKA KYKAREAALK AQSERKPGGE R
Mm.UBF_4        PKRPVSAMFI FSEEKRRQLQ EERP--ELSE SELTRLLARM WNDLTEKKKA KYKAREAALK AQSERKPGGE R
Hs.UBF_2A       PEKPKT-PQQ LWYTHEKKVY LK-VRPDATT KEVKDSLGKQ WSQLSDKKRL KWIHKALEQR KEYEEIMRDY I
Rn.UBF_2        PEKPKT-PQQ LWYTHEKKVY LK-VRPDATT KEVKDSLGKQ WSQLSDKKRL KWIHKALEQR KEYEEIMRDY I
Mm.UBF_2        PEKPKT-PQQ LWYTHEKKVY LK-VRPDATT KEVKDSLGKQ WSQLSDKKTL KWIHKALEQR KEYEEIMRDY I
Xl.UBF_2        PEKPKT-PQQ LWYNHERKVY LK-LHADAST KDIKDALGKQ WSQLPDKKRL KWIHKALEQR KQYEGVMREY M
Hs.UBF_3        -TKP---PPN SYSLYCAELM AN--MKDVPS TERMVLCSQQ WKLLSQKEKD AYHKKCDQKK KDYEVELLRF L
Mm.UBF_3        -TKP---PPN SYSLYCAELM AN--MKDVPS TERMVLCSQQ WKLLSQKEKD AYHKKCDQKK KDYEVELLRF L
Rn.UBF_3        -TKP---PPN SYSLYCAELM AN--MKDVPS TERMVLCSQQ WKLLSQKEKD AYHKKCDQKK KDYEVELLRF L
Hs.NOR90_2      -TKP---PPN SYSLYCAELM AN--MKDVPS TERMVLCSQQ WKLLSQKEKD AYHKKCDQKK KDYEVELLRF L
Xl.UBF_3        -TKP---PPN SYSMYCAELM AN--MKDVPS TERMVLCSQR WKLLSQKEKD AYHKKCEQRK KDYEVELMRF L
Hs.UBF_5        PKKP---PMN GYQKFSQELL SNGELNHLPL KERMVEIGSR WQRISQSQKE HYKKLAEEQQ KQYKVHLDLW V
Hs.NOR90_4      PKKP---PMN GYQKFSQELL SNGELNHLPL KERMVEIGSR WQRISQSQKE HYKKLAEEQQ KQYKVHLDLW V
Mm.UBF_5        PKKP---PMN GYQKFSQELL SNGELNHLPL KERMVEIGSR WQRISQSQKE HYKKLAEEQQ RQYKVHLDLW V
Rn.UBF_5        PKKP---PMN GYQKFSQELL SNGELNHLPL KERMVEIGSR WQRISQSQKE HYKKLAEEQQ RQYKVHLDLW V
Pa.FPR1         IPRPPNAYIL YRKDQQAALK AANP--GIPN NDISVMTGGM WKKESPEVRA EYQRRASEIK AKLMSAHPHY R
```

```
Nc.MATA1       IPRPPNAYIL YRKDHHREIR EQNP--GLHN NEISVIVGNM WRDEQPHIRE KYFNMSNEIK TRLLLENPDY R
Sp.MTMC        TPRPPNAFIL YRKEKHATLL KSNP--SINN SQVSKLVGEM WRNESKEVRM RYFKMSEFYK AQHQKMYPGY K
Hs.SRY         VKRPMNAFIV WSRDQRRKMA LENP--RMRN SEISKQLGYQ WKMLTEAEKW PFFQEAQKLQ AMHREKYPNY K
Gog.SRY        VKRPMNAFIV WSRDQRRKMA LENP--RMRN SEISKQLGYQ WKMLTEAEKW PFFQEAQKLQ AMHREKYPNY K
Pt.SRY         VKRPMNAFFV WSRDQRRKMA LENP--RMRN SEISKQLGYQ WKMLTEAEKW PFFQEAQKLQ AMHREKYPNY K
Hl.SRY         VKRPMNAFIV WSRDQRRKMA LENP--KMRN SEISKQLGYR WKMLTEAEKW PFFQEAQKLQ AMHREKYPNY K
Pp.SRY         VKRPMNAFIV WSRDQRRKMA LENP--KMRN SEISKQLGYQ WKMLTEAEKW PFFQEAQKLQ AMHREKYPNY K
Cj.SRY         VKRPMNAFIV WSRDQRRKMA VENP--QMRN SEISKRLGYQ WKLLTEAEKW PFFQEAQKLQ AMHREKYPNY K
Oa.SRY         VKRPMNAFIV WSRERRRKVA LENP--KLQN SEISKQLGYE WKRLTDAEKR PFFEEAQRLL AIHRDKYPGY K
Ch.SRY         VKRPMNAFIV WSRERRRKVA LENP--KLQN SEISKQLGYE WKRLTDAEKR PFFEEAQRLL AIHRDKYPGY K
Bt.SRY         VKRPMNAFIV WSRERRRKVA LENP--KMKN SDISKQLGYE WKRLTDAEKR PFFEEAQRLL AIHRDKYPGY K
Mm.SRYG        VKRPMNAFMV WSRGERHKLA QQNP--SMQN TEISKQLGCR WKSLTEAEKR PFFQEAQRLK TLHREKYPNY K
Mma.SRY        VKRPMNAFMV WSRGERHKLA QQNP--SMQN TEISKQLGCR WKSLTEAEKR PFFQEAQRLK TLHREKYPNY K
Ms.SRY         VKRPMNAFMV WSRGERHKLA QQNP--SMQN TEISKQLGCR WKSLTEAEKR PFFQEAQRLK TLHREKYPNY K
Mp.SRY         VKRPMNAFMV WSRGERHKLA QQNP--SMQN TEISKQLGCR WKSLTEAEKR PFFQEAQRLK TLHREKYPNY K
Mco.SRY        VKRPMNAFMV WSRGERHKLA QQNP--SMQN TEISKQLGCR WKSLTEDEKR PFFQEAQRLK TLHREKYPNY K
Mce.SRY        VKRPMNAFMV WSRGERHKLA QQNP--SMQN TEISKQLGCR WKSLTEDEKR PFFQEAQRLK TLHREKYPNY K
Mm.SRYA        VKRPMNAFMV WSRGERHKLA QQNP--SMQN TEISKQLGCR WKSLTEAEKR PFFQEAQRLK ILHREKYPNY K
Mca.SRY        IKRPMNAFMV WSRGERHKLA QQNP--SMQN TEISKQLGCR WKSLTEAEKR PFFQEAQRLK TLHREKYPNY K
Mh.SRY         VKRPMNAFMA WSRGERRKLA QQNP--SMQN TEISKQLGYR WKSLTEAEKR PFFQEAQRLK TLHREKYPNY K
Hs.SOX3        VKRPMNAFMV WSRGQRRKMA LENP--KMHN SEISKRLGAD WKLLTDAEKR PFIDEAKRLR AVHMKEYPDY K
Sm.SOX3        VKRPMNAFMV WSRGQRRKMA LENP--KMHN SEISKRLGAD WKLLTDAEKR PFIDEAKRLR AVHMKEYPDY K
Gg.SOX3        VKRPMNAFMV WSRGQRRKMA QENP--KMHN SEISKRLGAD WKLLSDAEKR PFIDEAKRLR AVHMKEYPDY K
Dr.SOX19       VKRPMNAFMV WSRGQRRKMA QENP--KMHN SEISKRLGAE WKLLTDAEKR PFIDEAKRLR ALHMKEYPDY K
Gg.SOX2        VKRPMNAFMV WSRGQRRKMA QENP--KMHN SEISKRLGAE WKLLSEAEKR PFIDEAKRLR ALHMKEHPDY K
Hs.SOX         VKRPMNAFMV WSRGQRRKMA QENP--KMHN SEISKRLGAE WKLLSETEKR PFIDEAKRLR ALHMKEHPDY K
Dm.SOX70D      IKRPMNAFMV WSRLQRRQIA KDNP--KMHN SEISKRLGAE WKLLAESEKR PFIDEAKRLR ALHMKEHPDY K
Hs.SOX20       VKRPMNAFMV WSSAQRRQMA QQNP--KMHN SEISKRLGAQ WKLLDEDEKR PFVEEAKRLR ARHLRDYPDY K
Sm.SRY         VKRPMNAFMV WSQTQRRKVA LQNP--KMHN SEISKQLGVT WKLLSDSEKR PFIDEAKRLR DKHKQVS-DY K
Hs.SOX9        VKRPMNAFMV WAQAARRKLA DQYP--HLHN AELSKTLGKL WRLLNESEKR PFVEEAERLR VQHKKDHPDY K
Mm.SOX9        VKRPMNAFMV WAQAARRKLA DQYP--HLHN AELSKTLGKL WRLLNESEKR PFVEEAERLR VQHKKDHPDY K
Gg.SOX9        VKRPMNAFMV WAQAARRKLA DQYP--HLHN AELSKTLGKL WRLLNESEKR PFVEEAERLR VQHKKDHPDY K
Mm.SOX21       VKRPMNAFMV WAQAARRKLA DQYP--HLHN AELSKTLGKL WRLLNESDKR PFIEEAERLR MQHKKDHPDY K
Mm.SOX18       IRRPMNAFMV WAKDERKRLA QQNP--DLHN AVLSKMLGKA WKELNTAEKR PFVEEAERLR VQHLRDHPNY K
Hs.SOX4        IKRPMNAFMV WSQIERRKIM EQSP--DMHN AEISKRLGKR WKLLKDSDKI PFIREAERLR LKHMADYPDY K
Mm.SOX4        IKRPMNAFMV WSQIERRKIM EQSP--DMHN AEISKRLGKR WKLLKDSDKI PFIQEAERLR LKHMADYPDY K
Hs.SOX11       IKRPMNAFMV WSKIERRKIM EQSP--DMHN AEISKRLGKR WKMLKDSEKI PFIREAERLR LKHMADYPDY K
Gg.SOX11       IKRPMNAFMV WSKIERRKIM EQSP--DMHN AEISKRLGKR WKMLKDSEKI PFIREAERLR LKHMADYPDY K
Hs.SOX22       IKRPMNAFMV WSQHERRKIM DQWP--DMHN AEISKRLGRR WQLLQDSEKI PFVREAERLR LKHMADYPDY K
Dm.SOX14       IKRPMNAFMV WSQMERRKIC ERTP--DLHN AEISKELGRR WQLLSKDDKQ PYIIEAEKLR KLHMIEYPNY K
Mm.SOX6        IKRPMNAFMV WAKDERRKIL QAFP--DMHN SNISKILGSR WKSMSNQEKQ PYYEEQARLS KIHLEKYPNY K
Mm.SOXLZ       IKRPMNAFMV WAKDERRKIL QAFP--DMHN SNISKILGSR WKSMSNQEKQ PYYEEQARLS KIHLEKYPNY K
Om.SOXLZ       IKRPMNAFMV WAKDERRKIL QAFP--DMHN SNISKILGSR WKSMTNQEKQ PFYEEQARLS KIHLEKYPNY K
Hs.SOX5        IKRPMNAFMV WAKDERRKIL QAFP--DMHN SNISKILGSR WKAMTNLEKQ PYYEEQARLS KQHLEKYPDY K
Mm.SOX5        IKRPMNAFMV WAKDERRKIL QAFP--DMHN SNISKILGSR WKAMTNLEKQ PYYEEQARLS KQHLEKYPDY K
Hs.TCF4        IKKPLNAFML YMKEMRAKVV AECT--LKES AAINQILGRR WHALSREEQA KYYELARKER QLHMQLYPGW S
Mm.LEF-1       IKKPLNAFML YMKEMRANVV AECT--LKES AAINQILGRR WHALSREEQA KYYELARKER QLHMQLYPGW S
Hs.TCF1        IKKPLNAFML YMKEMRAKVI AECT--LKES AAINQILGRR WHALSREEQA KYYELARKER QLHMQLYPGW S
Hs.TCF3        VKKPLNAFML YMKEMRAKVV AECT--LKES AAINQILGRK WHNLSREEQA KYYELARKER QLHSQLYPTW S
Rn.HBP1        CKRPMNAFML FAKKYRVEYT QMYP--GKDN RAISVILGDR WKKMKNEERR MYTLEAKALA EEQKRLNPDC W
CIV.HMG        PKRNKSSYLF FCQEIRPSIV AEMP--DIKP NQVMVHLGKK WSELPLEDRK KYDVMAVEDR KRYLASKEAN K
Sc.MTHMG_1     PKRPTSAYFL YLQDHRSQFV KENP--TLRP AEISKIAGEK WQNLEADMKE KYISERKKLY SEYQKAKKEF D
Sc.ROX1        -PRPKNAFIL FRQHYHRILI DEWTAQIPHN SNISKIIGTK WKGLQPEDKA HWENLAEKEK LEHERKYPEY K
Hs.PMS1        IKKPMSASAL FVQDHRPQFL IENP--KTSL EDATLQIEEL WKTLSEEEKL KYEEKATKDL ERYNSQMKRA I
Sc.IXR1_1      PKRPSSAYFL FSMSIRNELL QQFP--EAKV PELSKLASAR WKELTDDQKK PFYEEFRTNW EKYRVVRDAY E
Pf.HMGB4       PKAPPSSYLI FCNYERENAK NTLLQKTIRI TDIQKELSNK WKNLPEDERK KYEEQAQILK SKYNEELLEW K
Tt.MLH         PKKPIGSFFR FLEENRQKYA AKHK--DLTN AKILKIMSED FNNLPQKEVK VYEDAYQKEY AQYLVEFKKW N
Sc.YD9395      PKKPLTVFFA YSAYVRQELR EDRQKAPLSS TEITQEISKK WKELSDNEKE KWKQAYNVEL ENYQREKSKY L
Sc.IXR1_2      PKRPSGPFIQ FTQEIRPTVV KENP--DKGL IEITKIIGER WRELDPAKKA EYTETYKKRL KEWESCYPDE N
Pf.HMGB3_2     -KRKFTAFSI FAREKRKEYK EKNIDM-LTL AQQNSHVSKL WKQLTAEEKN KYKVLSNTVN ATIAKAYSKS E
Pf.HMGB3_1     AIRGITSFTL FAREKRKELL DQKIYL-SSL TEQTSAVAKI WNNLSDEQKK EWAVKASKIN EENFLLQKKK K
```

I.4 - Numéros d'accession des séquences utilisées pour la phylogénie des facteurs NF-YB et NF-YC

Les numéros d'accession sont ceux de la base de données UniProt. Les noms des facteurs sont codés de la manière suivante : « abréviation_organisme.nom-facteur ».

P25208, Hs.NF-YB ; O73744, Xl.NF-YB ; Q8ST61, Dm.NF-YB ; O23310, At.NF-YB-3 ; O17286, Ce.NF-YB ; P36611, Sp.HAP3 ; P13434, Sc.HAP3 ; Q13952, Hs.NF-YC ; O73745, Xl.NF-YC ; Q9W3V9, Dm.NF-YC ; Q9ZVL3, At.NF-YC-9 ; O17072, Ce.NF-YC ; P79007, Sp.HAP5 ; Q02516, Sc.HAP5.

I.5 - Alignement utilisé pour la phylogénie des facteurs NF-YB et NF-YC

Les domaines 'CBFD_NFYB_HMF' identifiés par MotifScan (Pfam : PF00808) sont indiqués en majuscules.

```
Hs.NFYB     skesfreqdI YLPIANVARI MKNAIPQTG- -KIAKDAKEC VQECVSEFIS
Xl.NFYB     skdsfreqdI YLPIANVARI MKNAVPQTG- -KIAKDAKEC VQECVSEFIS
Dm.NFYB     ggimlreqdR FLPICNIIKI MKVPVPQNG- -KIAKDAREC IQECVSEFIS
At.NFYB-3   gnastreqdR FLPIANVSRI MKKALPANA- -KISKDAKET VQECVSEFIS
Ce.NFYB     ksqvlldqeR FLPIANVVRI MKTQMDPQA- -KLAKDAKEC AQECVSEFIS
Sp.HAP3     msadgldytn LLPIANVARI MKSALPENA- -KISKEAKDC VQDCVSEFIS
Sc.HAP3     qistlreqdR WLPINNVARL MKNTLPPSA- -KVSKDAKEC MQECVSELIS
PF11_0477   kkgskcdseT LLPIANISRI MKRILPGSA- -KVAKESKDI IRECVTEFIQ
Hs.NFYC     nltvkdfrvQ ELPLARIKKI MKLDEDVK-- -MISAEAPVL FAKAAQIFIT
Xl.NFYC     nltvkdfrvQ DLPLARIKKI MKLDEDVK-- -MISAEAPVL FAKAAQIFIT
Dm.NFYC     sigqvdakhQ VLPLARIKKI MKLDENAK-- -MIAGEAPLL FAKACEYFIQ
At.NFYC-9   iekttdfknH SLPLARIKKI MKADEDVR-- -MISAEAPVV FARACEMFIL
Sc.NFYC     sehqddfksH SLPFARIRKV MKTDEDVK-- -MISAEAPII FAKACEIFIT
Sp.NFYC     ehddqavktL HLPLARIKKV MKTDDDVKNK -MISAEAPFL FAKGSEIFIA
Ce.NFYC     edmlnksknM SVPMARVKKI MRIDDDVRNF -MIASDAPIF MAQAAEFFIE
PF14_0374   nmspedlkiH NLPISRIKKI MKEDDEIKSN QMVSADTPVL LAKACELFIM
PF13_0043   qtnkiketlf glstgiiqka innnvdlrn- yRMRKEALET LGKCLSMFIL

            FITSEASERC HQEKRKTING EDILFAmstl g-fdsyvepl klylqkf
            FITSEASERC HQEKRKTING EDILFAmsrl g-fdsyvepl klylqkf
            FISSEAIERS VAENRKTVNG DDLLVAfsnl g-fdnyvepl siylqky
            FITGEASDKC QREKRKTING DDLLWAmttl g-fedyvepl kvylqky
            FIASEAAEIC NITKRKTITA DDLLTAmeat g-fdnyaepm riflqky
            FVTGEASEQC TQEKRKTITG EDVLLAlntl g-fenyaevl kisltky
            FVTSEASDRC AADKRKTING EDILISlhal g-fenyaevl kiylaky
            FLTSEASDRC TREKRKTING EDILYSMEkl g-fndyiepl teylnkw
            ELTLRAWIHT EDNKRRTLQR NDIAMAItkf dqfdflidiv prdelkp
            ELTLRAWIHT EDNKRRTLQR NDIAMAItkf dqfdflidiv prdelkp
            ELTMHAWVHT EESRRRTLQR SDIAQAIany dqfdflidiv preeikp
            ELTLRSWNHT EENKRRTLQK NDIAAAVtrt difdflvdiv predlrd
            ELTMRAWCVA ERNKRRTLQK ADIAEALqks dmfdflidvv prrplpq
            ELTMRAWLHA KKNQRRTLQR SDIANAVsks emydflidii skdnnns
            EMTAMGWQYV SEARRRILQK ADIASAVqks dqfdflidfl ppktvpt
            ELTSNAWKYT EEGKRRTLQR QDVVSAAckk dtfdflidli pledrmk
            YITDGAMEYC ENEKRSTILV RDILNSLdds lfldihdelk rqltiqe
```

II - Modélisation par homologie

II.1 - Les facteurs PfHMGB1 & PfHMGB2

a. Alignements

(i) PfHMGB1.ali

```
>P1;1J5N
structureX:1J5N: 1:A:93: :ScNHP6A : : :
MVTPREPKKRTTRKKKDPNAPKRALSAYMFFANENRDIVRSENPDI--TF
GQVGKKLGEKWKALTPEEKQPYEAKAQADKKRYESEKELYNATLA--
*
>P1;PfHMGB1
sequence:PfHMGB1: 1: :97 : : : : :
MKNTGKEVRKRRKNKKDPHAPKRSLSAYMFFAKEKRAEIISKQPELSKD
VATVGKMIGEAWNKLGEKEKAPFEKKAQEDKLRYEKEKAEYANMKMKA
*
```

(ii) PfHMGB2.ali

```
>P1;1J5N
structureX:1J5N: 1:A:93: :ScNHP6A : : :
---MVTPREPKKRTTRKKKDPNAPKRALSAYMFFANENRDIVRSENPDI-
-TFGQVGKKLGEKWKALTPEEKQPYEAKAQADKKRYESEKELYNATLA-
*
>P1;PfHMGB2
sequence:PfHMGB2: 1: :99 : : : : :
MASKSQKKVLKKQNKKKKKDPLAPKRALSAYMFYVKDKRLEIIKEKPELA
KDVAQVGKLIGEAWGQLSPAQKAPYEKKAQLDKVRYSKEIEEYRKKNQE
*
```

b. Fonctions objectives

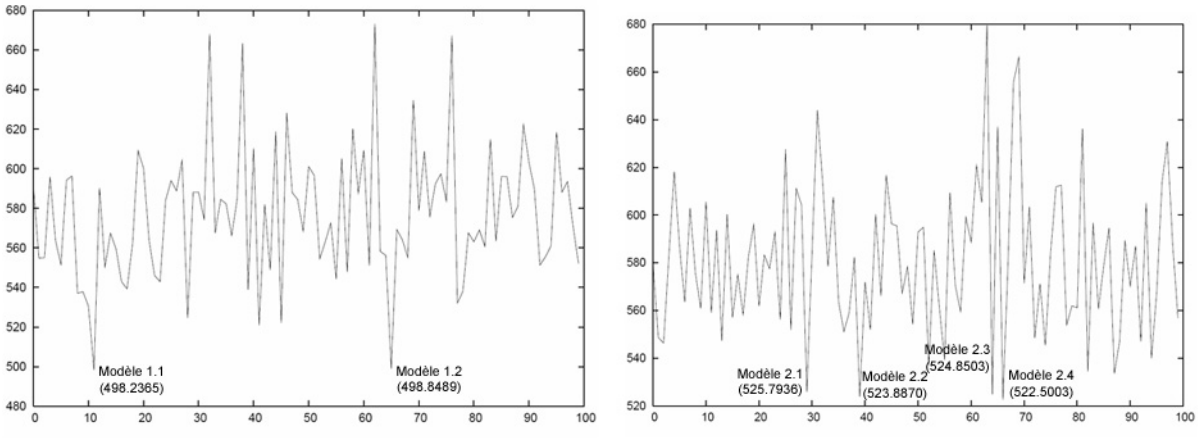


Figure 64. Fonctions objectives des 100 modèles de PfHMGB1 & PfHMGB2.
En abscisse, les 100 modèles ; en ordonnée, les valeurs des fonctions objectives. Les modèles retenus sont indiqués sur les graphiques avec leur fonction objective respective entre parenthèses.

c. Qualité des structures modèles

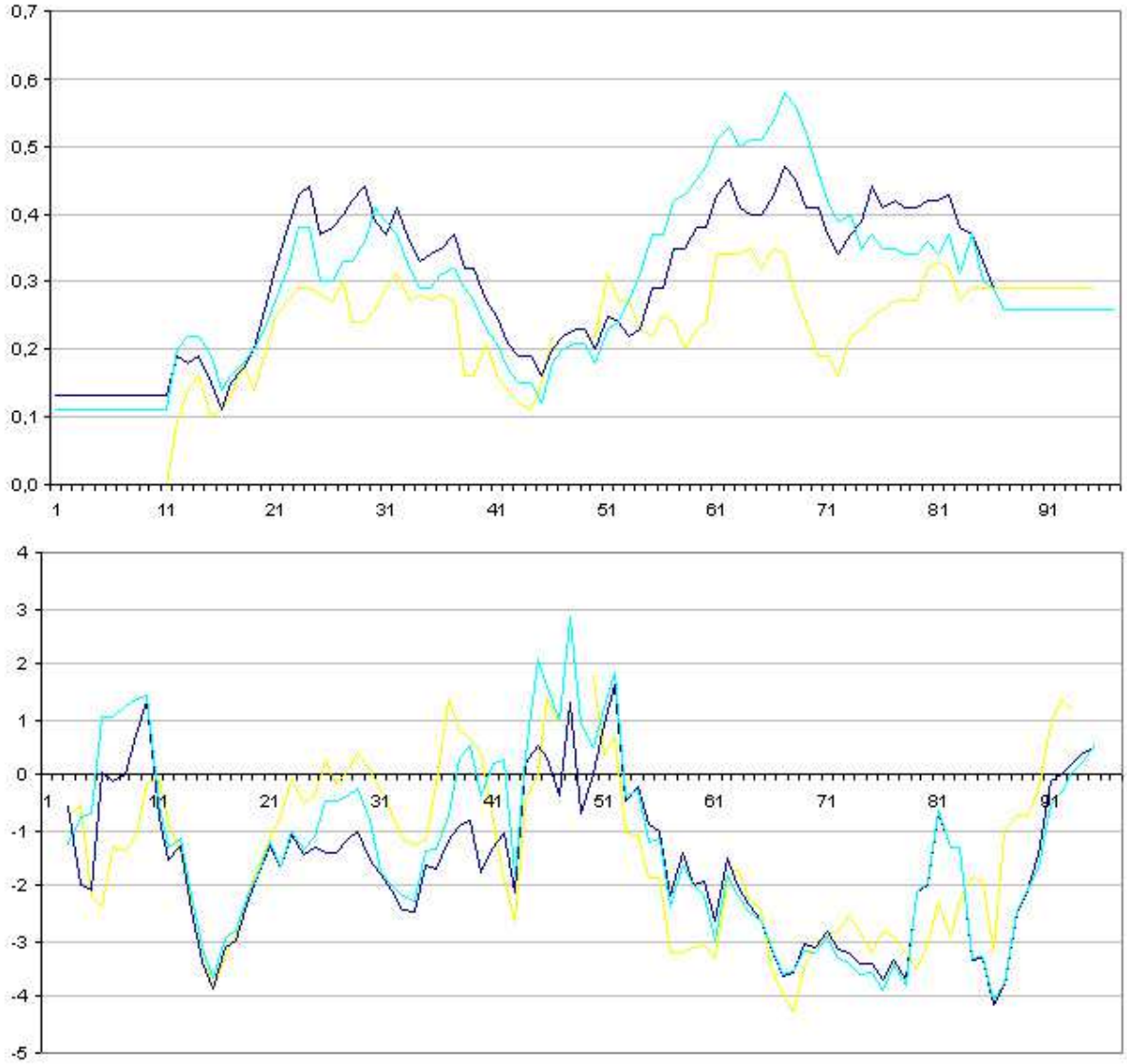


Figure 65. Analyse de la structure support 1J5N et des deux structures modèles de PfHMGB1 par Verify3D (en haut) et ProSa2003 (en bas).
Les courbes sont les suivantes : en jaune, la structure support 1J5N ; en bleu, le modèle 1.1 ; en cyan, le modèle 1.2. Les trois courbes sont superposées en respectant l'alignement utilisé pour la simulation avec Modeller. En abscisse sont indiquées les positions des sites de l'alignement. **(Haut)** Verify3D. En ordonnée se trouve le score moyen assigné à chaque acide aminé : plus le score est élevé, meilleur est le modèle. **(Bas)** ProSa2003. En ordonnée se trouve le score assigné à chaque acide aminé. Plus le score est faible, meilleur est le modèle.

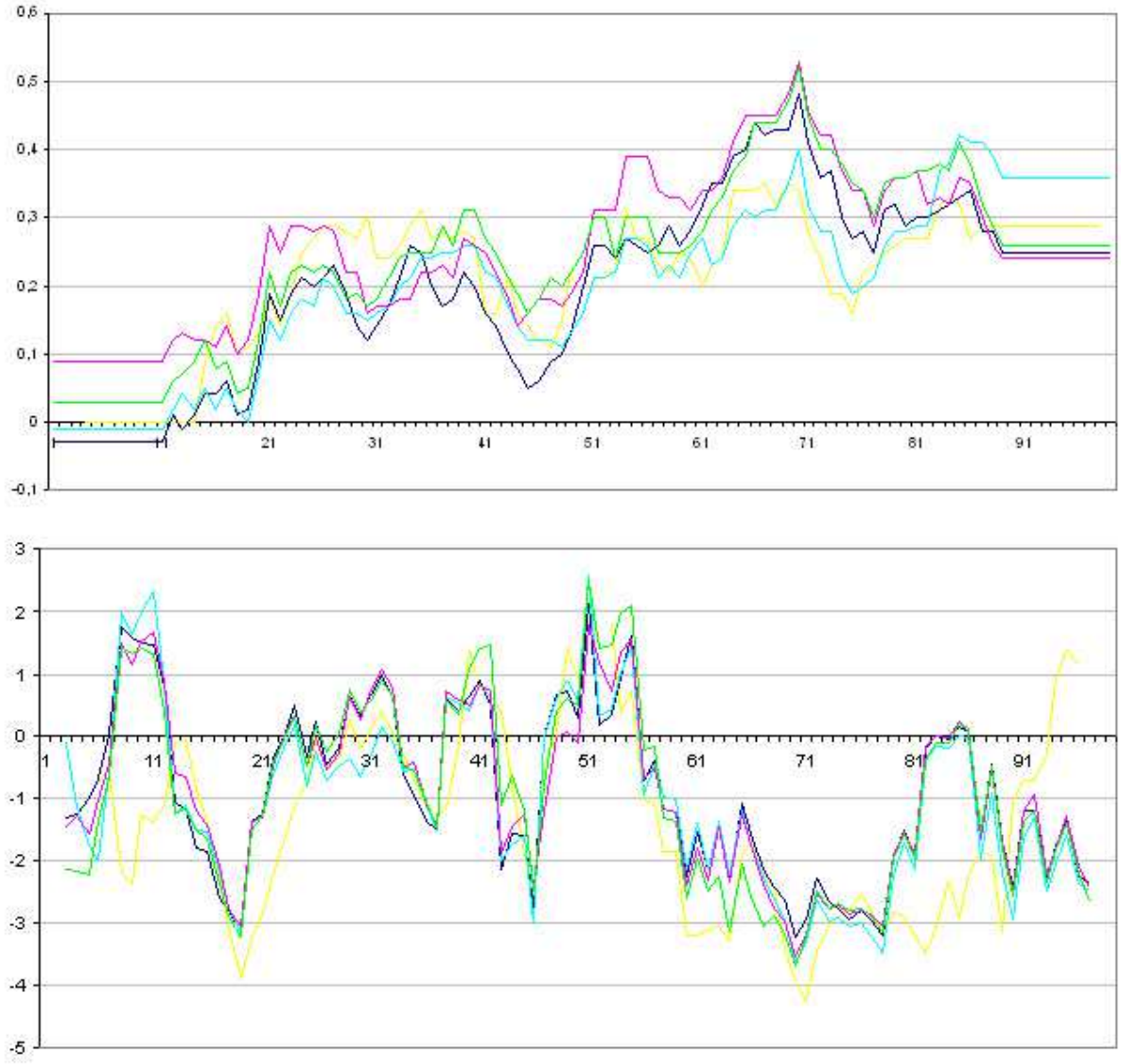


Figure 66. Analyse de la structure support 1J5N et des quatre structures modèles de PfHMGB2 par Verify3D (en haut) et ProSa2003 (en bas).

Les courbes sont les suivantes : en jaune, la structure support 1J5N ; en bleu, le modèle 2.1 ; en cyan, le modèle 2.2 ; en rose, le modèle 2.3 et en vert, le modèle 2.4. Les cinq courbes sont superposées en respectant l'alignement utilisé pour la simulation avec Modeller. En abscisse sont indiquées les positions des sites de l'alignement. **(Haut)** Verify3D. En ordonnée se trouve le score moyen assigné à chaque acide aminé : plus le score est élevé, meilleur est le modèle. **(Bas)** ProSa2003. En ordonnée se trouve le score assigné à chaque acide aminé. Plus le score est faible, meilleur est le modèle.

II.2 - Modélisation du facteur PfMyb1

a. Alignements

Dans le cas de la première méthode de modélisation

(i) PfMyb1.ali

```
>P1;1H88
structureX:1H88: 41:C:175: :c-Myb : : :
TRWTREEDEKLKKLVEQNGTDDWKVIANYLPN-RTDVQCQHRWQK--VLN
PELIKGPWTKEEDQRVIKLVQKYGPKRWSVIAKHLKG-RIGKQCRERWHN
HL----NPE-VKKT--SWTEEEDRIIYQAHKRLGN-RWAEIAKLLPG
*
>P1;PfMyb1
sequence:PfMyb1: 221: :365 : : : : :
FSYTKN-DKNIEHNFLYFSETFWNEVSEKLSNNQNAKECQKMWLYYGCFE
DDKQK-KWTKDEVDKLLCLSKKYEQRNWKCIARELNTNRSPLSCFEQYIK
INKLYENKEKVKLERIAFNVLEDIQLQILVSIIGDKNWAEVKKHMES
*
```

Dans le cas de la deuxième méthode de modélisation

(i) PfMyb1.ali

```
>P1;1H88
structureX:1H88: 41:C:138: :c-Myb : : :
TRWTREEDEKLKKLVEQNGTDDWKVIANYLPN-RTDVQCQHRWQK--VLN
PELIKGPWTKEEDQRVIKLVQKYGPKRWSVIAKHLKG-RIGKQCRERWHNHL
*
>P1;PfMyb1
sequence:PfMyb1: 221: :320 : : : : :
FSYTKN-DKNIEHNFLYFSETFWNEVSEKLSNNQNAKECQKMWLYYGCFE
DDKQK-KWTKDEVDKLLCLSKKYEQRNWKCIARELNTNRSPLSCFEQYIKIN
*
```

ou

```
>P1;1H88
structureX:1H88: 42:C:123: :c-Myb : : :
RWTREEDEKLKKLVEQNGTDDWKVIANYLP-NRTDVQCQHRWQKV----L
NPELIKGP---WTKEEDQRVIKLVQKYGPKRWSVIAKHLK
*
>P1;PfMyb1
sequence:PfMyb1: 275: :364 : : : : :
KWTKDEVDKLLCLSKKYEQRNWKCIARELNTNRSPLSCFEQYIKINKLYE
NKEKVKLERIAFNVLEDIQLQILVSIIGDKNWAEVKKHME
*
```

b. Fonctions objectives

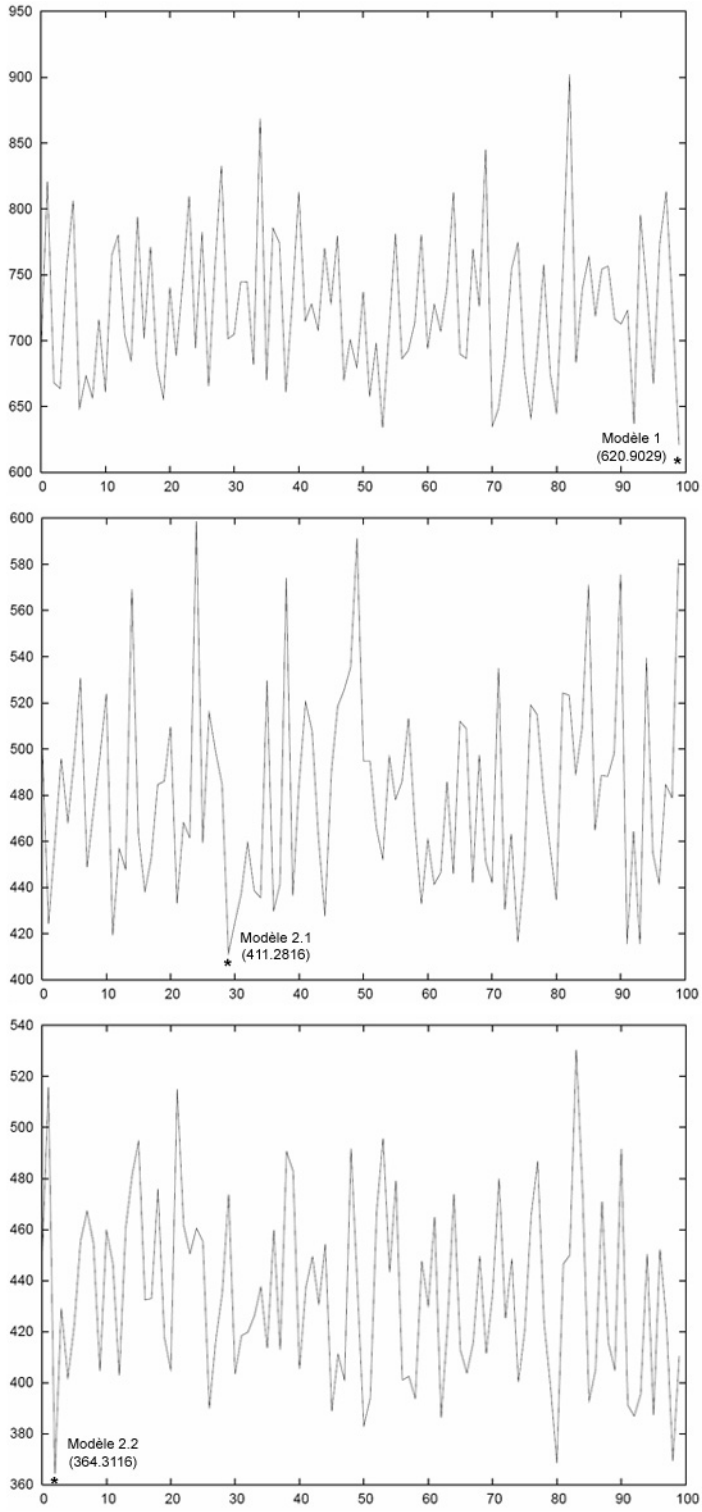


Figure 67. Fonctions objectives des 100 structures modèles obtenues pour chaque simulation.
En abscisse, les 100 modèles ; en ordonnée, les valeurs des fonctions objectives. Le modèle sélectionné dans chaque simulation est indiqué sur le diagramme correspondant par un astérisque avec sa fonction objective associée, entre parenthèses.

c. Qualité des structures modèles

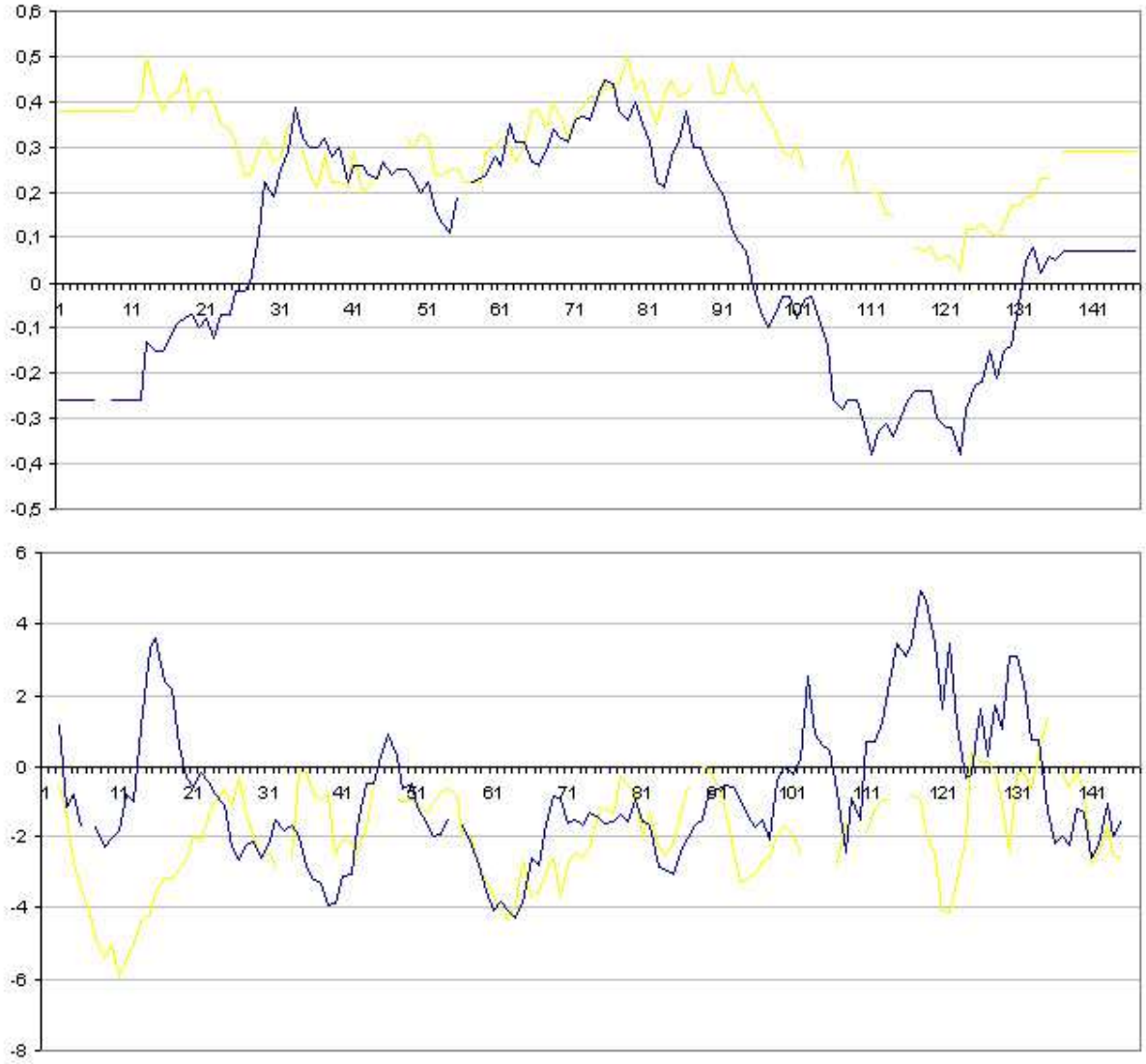


Figure 68. Analyse de la structure support 1H88 et du modèle 1 du domaine de liaison à l'ADN de PfMyb1 par Verify3D (en haut) et ProSa2003 (en bas).
La courbe jaune correspond à la structure support 1H88, la courbe bleue à la structure modèle de PfMyb1. Les deux courbes sont superposées en respectant l'alignement utilisé pour la simulation avec Modeller ; en abscisse sont indiquées les positions des sites de l'alignement. **(Haut)** Verify3D. En ordonnée se trouve le score moyen assigné à chaque acide aminé : plus le score est élevé, meilleur est le modèle. **(Bas)** ProSa2003. En abscisse les séquences protéiques, en ordonnée le score assigné à chaque acide aminé. Plus le score est faible, meilleur est le modèle.

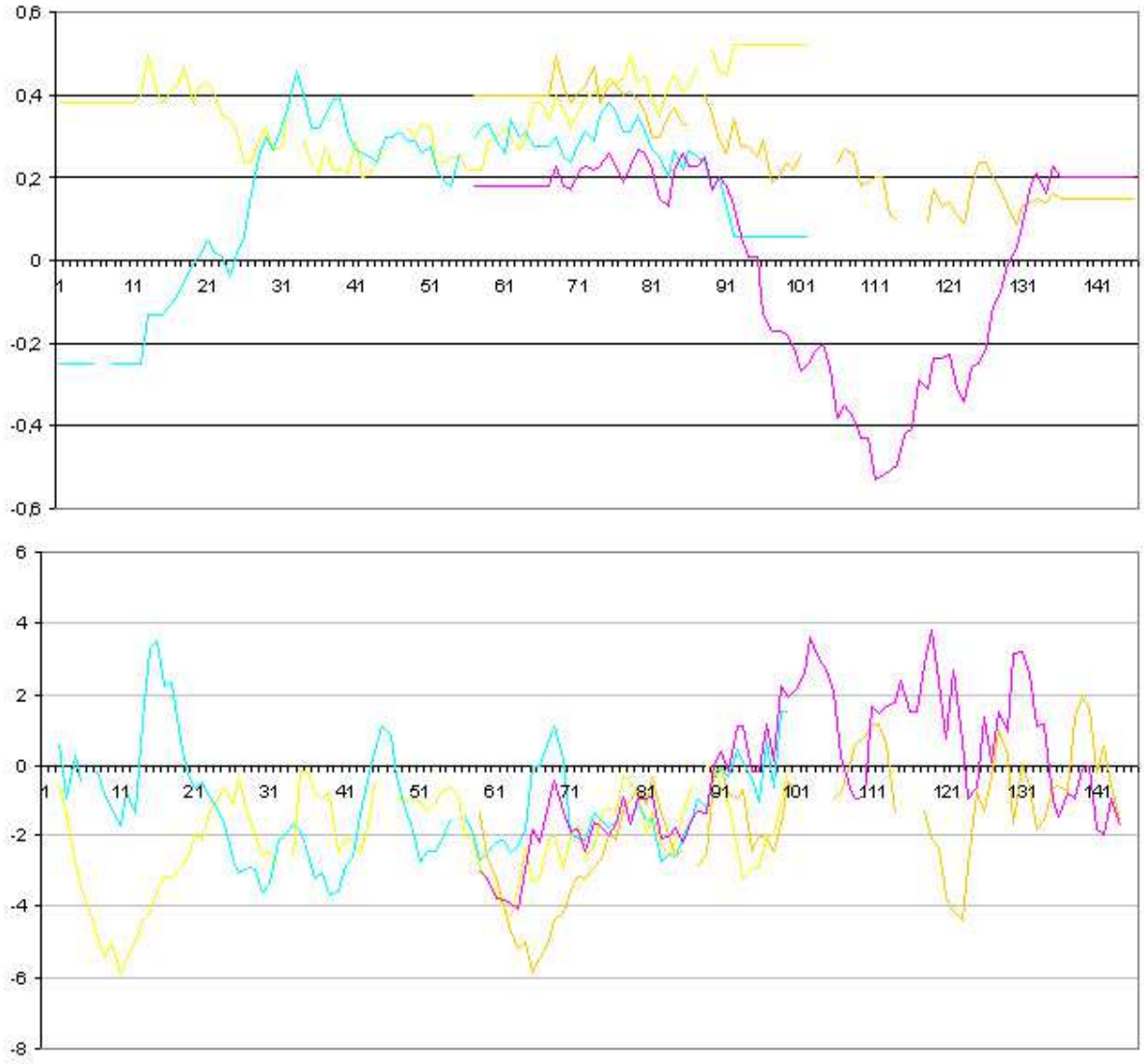


Figure 69. Analyse de la structure support 1H88 et de la structure modèle du début et de la fin du domaine de liaison de PfMyb1 par Verify3D (en haut) et ProSa2003 (en bas).
Les structures supports utilisées pour modéliser le début et la fin du domaine de liaison de PfMyb1 sont en jaune clair et jaune foncé. Les structures modèles du début et de la fin du domaine de liaison de PfMyb1 (modèles 2.1 et 2.2) sont respectivement en cyan et en magenta. Les courbes sont superposées en respectant l'alignement utilisé pour la simulation avec Modeller et la position des séquences les unes par rapport aux autres ; en abscisse sont indiquées les positions des sites de l'alignement. **(Haut)** Verify3D. En ordonnée se trouve le score moyen assigné à chaque acide aminé : plus le score est élevé, meilleur est le modèle. **(Bas)** ProSa2003. En abscisse les séquences protéiques, en ordonnée le score assigné à chaque acide aminé. Plus le score est faible, meilleur est le modèle.

III - Protéines liées à la régulation de l'expression des gènes chez *P. falciparum*

III.1 - Protéines annotées par Robert Coulson et ses collaborateurs [68]

N° accession	**Fonctions putatives**
PFC0805w	RNA polymerase II::Rpb-1
PFB0715w	RNA polymerase II::Rpb-2
PFI1130c	RNA polymerase II::Rpb-3
PFB0245c	RNA polymerase II::Rpb-4
PF13_0341	RNA polymerase I/II/III::Rpb-5 [ABC27]
PFC0155c	RNA polymerase I/II/III::Rpb-6 [ABC23]
PF10_0269	RNA polymerase II::Rpb-7
PFL0665c	RNA polymerase I/II/III::Rpb-8 [ABC14.5]
PFA0505c	RNA polymerase II::Rpb-9
PF07_0027	RNA polymerase II::Rpb-10 [ABC10-beta]
PF13_0023	RNA polymerase II::Rpb-11
MAL13P1.213	RNA polymerase II::Rpb-12 [ABC10-alpha]
PFE0465c	RNA polymerase I::A190
PFF11_0358	RNA polymerase I::A135
PFD0360w	RNA polymerase I::A12.2
PF11_0445	RNA polymerase I/III::AC40
PF14_0150	RNA polymerase I/III::AC19
PF13_0150	RNA polymerase III::C160
PFL0330c	RNA polymerase III::C128
PF14_0207	RNA polymerase III::C34
PF11_0058	RNA polymerase III::C25
PFB0290c	RNA polymerase III::C11
PF11_0264	mitochondrial RNA polymerase
PF14_0695	apicoplast RNA polymerase:: alpha subunit
PF13_0040	apicoplast RNA polymerase:: alpha subunit
PF10_0367	apicoplast RNA polymerase:: beta, beta' & beta'' subunits
PF11_0315	apicoplast RNA polymerase:: beta, beta' & beta'' subunits
PFE0635c	apicoplast RNA polymerase:: beta, beta' & beta'' subunits
PFI0330c	apicoplast RNA polymerase:: beta, beta' & beta'' subunits
PFA0525w	TFIIB::TFIIB
PFE0415w	TFIIB::TFIIB-like
PFE0305w	TFIID::TBP
PF14_0267	TFIID::TBP-like factor
MAL7P1.86	TFIIE::TFIIE-alpha
PF10_0369	TFIIH::XPB
PFI1650w	TFIIH::XPD
PFL2125c	TFIIH::p52
MAL13P1.76	TFIIH::p44
PF13_0279	TFIIH::p34
PFE0610c	TFIIH::MAT1

PF13_0022	TFIIH::cyclin K
PF14_0469	TFIIIB::70/90 kDa subunits
PFL0520c	TFIIIC::63 kDa subunit
PF10_0293	Transcription-elongation::SPT4
MAL6P1.111	Transcription-elongation::SPT5
PF14_0059	Transcription-elongation::SPT6
PF07_0057	Transcription elongation factor TFIIS
PFE0870w	Transcription elongation factor CDC68
PFA0355w	CCR4-NOT complex::CCR4
PFE0980c	CCR4-NOT complex::CCR4
PFC0850c	CCR4-NOT complex::CCR4
MAL8P1.104	CCR4-NOT complex::CAF1
PF11_0049	CCR4-NOT complex::NOT1
PF14_0170	CCR4-NOT complex::NOT1
PF11_0297	CCR4-NOT complex::NOT2
PF10_0062	CCR4-NOT complex::NOT3/NOT5
PFL1705w	CCR4-NOT complex::NOT4
PF11_0477	CCAAT box-binding complex::CBF-A/NF-Y-B
PF13_0043	CCAAT box-binding complex::CBF-B/NF-Y-A
PF14_0374	CCAAT box-binding complex::CBF-C/NF-Y-C
PF08_0034	Histone acetyltransferase GCN5
PF10_0143	Transcriptional adaptator ADA2
PF11_0053	Nucleosome remodelling::SNF2L
MAL6P1,183	Nucleosome remodelling::ISWI
PFI1260c	Histone deacetylase RPD3
PF10_0078	Histone deacetylase
PF14_0690	Histone deacetylase
MAL7P1.37	SIN3-repressing complex::p18
MAL6P1.269	Transcription co-activator ALY
PF11_0293	Transcription co-activator MBF1
PFC0175w	Transcription co-activator TRIAD2
PF14_0489	Sir2-like protein
PF07_0083	TBP-binding protein
PF14_0393	Single-strand recognition protein
MAL6P1.142	DNA methyltransferase 1-associated protein 1 DNMAP1
PFB0875c	Chromatin-binding protein (SKI/SNW family) SkiP
PF14_0241	General transcription factor BTF3
PFL2390c	Pleiotropic regulatory protein THO2
PF14_0718	RNA polymerase II-associated factor SOH1
PF14_0649	Transcription factor C1
MAL8P1.131	Transcription factor GAS41
PF07_0123	Transcriptional regulator SPO8
PF13_0293	SET-domain transcriptional regulator
PFE0090w	Histone transcription regulator
PFI1470c	Histone transcription regulator
PF14_0492	Calcium-dependent transcriptional repressor DREAM
MAL8P1.111	Transcriptional repressor RPH1

III.2 - Protéines annotées par nos soins

Numéros d'accession	**Annotations**	**Domaines caractéristiques**	**Superclasses**
MAL13P1.97	hypothetical protein	bZip	Domaines basiques
PF11_0433	hypothetical protein	bZip	Domaines basiques
PFB0615c	hypothetical protein	bZip	Domaines basiques
PFF0140c	hypothetical protein	bZip	Domaines basiques
PFF0375c	hypothetical protein	bZip	Domaines basiques
PFC0365w	hypothetical protein	bZip	Domaines basiques
PFL2185w	hypothetical protein	bZip	Domaines basiques
MAL13P1.140	hypothetical protein	bZip	Domaines basiques
MAL7P1.87	hypothetical protein	bZip	Domaines basiques
PF13_0047	hypothetical protein	bZip	Domaines basiques
PFA0550w	hypothetical protein	bZip	Domaines basiques
PFC0235w	hypothetical protein	bZip	Domaines basiques
PFC0760c	hypothetical protein	bZip	Domaines basiques
PFD0330w	hypothetical protein	bZip	Domaines basiques
PFF1285w	hypothetical protein	bZip	Domaines basiques
PFI0175w	hypothetical protein	bZip	Domaines basiques
PFL1930w	hypothetical protein	bZip	Domaines basiques
MAL13P1.37	hypothetical protein	B_box	Doigts de zinc
PF14_0383	Constans	B_box	Doigts de zinc
PFC0345w	hypothetical protein	B-box, C2H2	Doigts de zinc
PF10_0091	hypothetical protein	C2H2	Doigts de zinc
PF14_0479	hypothetical protein	C2H2	Doigts de zinc
PF14_0559	hypothetical protein	C2H2	Doigts de zinc
PF14_0643	hypothetical protein	C2H2	Doigts de zinc
PF14_0657	hypothetical protein	C2H2	Doigts de zinc
PFD0375w	hypothetical protein	C2H2	Doigts de zinc
PFL0455c	hypothetical protein	C2H2	Doigts de zinc
PFL2075c	hypothetical protein	C2H2	Doigts de zinc
PF14_0612	hypothetical protein	C2H2	Doigts de zinc
PF13_0278	hypothetical protein	C2H2	Doigts de zinc
PFB0725c	hypothetical protein	DHHC	Doigts de zinc
PFB0140w	hypothetical protein	DHHC	Doigts de zinc
MAL13P1.117	hypothetical protein	DHHC	Doigts de zinc
MAL7P1.68	hypothetical protein	DHHC	Doigts de zinc
PF10_0273	hypothetical protein	DHHC	Doigts de zinc
PF11_0167	hypothetical protein	DHHC	Doigts de zinc
PFE1415w	cell cycle regulator	DHHC	Doigts de zinc
PFF0485c	hypothetical protein	DHHC	Doigts de zinc
MAL13P1.126	hypothetical protein	DHHC	Doigts de zinc
PFI1580c	DHHC-type zinc finger protein	DHHC	Doigts de zinc
PF11_0217	hypothetical protein	DHHC	Doigts de zinc
PFI0590c	hypothetical protein	HIT	Doigts de zinc
PFF0105w	MYND finger domain protein	MYND	Doigts de zinc
PF07_0124	hypothetical protein	MYND	Doigts de zinc

PF13_0293	hypothetical protein	MYND	Doigts de zinc
PFF0350w	MYND finger protein	MYND	Doigts de zinc
MAL13P1.216	DNA helicase	RING	Doigts de zinc
PFD0765w	RING finger protein	RING	Doigts de zinc
PFC0690c	hypothetical protein	RING, C2H2	Doigts de zinc
PFI1255w	hypothetical protein	Yippee	Doigts de zinc
PF11_0241	hypothetical protein	Myb	Hélice-Tour-Hélice
PFL0815w	DNA-binding chaperone	Myb	Hélice-Tour-Hélice
PFL1215c	hypothetical protein	Myb	Hélice-Tour-Hélice
PF10_0327	PfMyb2	Myb	Hélice-Tour-Hélice
PF13_0088	PfMyb1	Myb	Hélice-Tour-Hélice
PFF0720c	hypothetical protein	Myb	Hélice-Tour-Hélice
PFF1385c	hypothetical protein	Myb	Hélice-Tour-Hélice
PFA0470c	cold shock protein	Cold_shock	Architecture bêta
MAL13P1.290	PfHMGB4	HMG-box	Architecture bêta
MAL8P1.72	PfHMGB2	HMG-box	Architecture bêta
PFL0145c	PfHMGB1	HMG-box	Architecture bêta

Numéros d'accession	**Annotations**	**Domaines caractéristiques**	**Fonctions**
PFL0465c	TFIIIA	C2H2	Facteur général
PFL0290w	TFIIIIb'	HMG-box, Myb	Facteur général
PF08_0037	hypothetical protein	MED7 Mediator	Mediator
PF11_0192	hypothetical protein	moz acetyl	Modification de la chromatine
PF13_0152	transcriptional regulatory	SIR2	Modification de la chromatine
PF14_0489	hypothetical protein	SIR2	Modification de la chromatine
PFD0190w	hypothetical protein	SET	Modification de la chromatine
PFL0690c	hypothetical protein	SET	Modification de la chromatine
PF11_0392	hypothetical protein	Bromodomain	Remodelage de la chromatine
PFF1440w	SET-domain protein	PHD	Remodelage de la chromatine
MAL13P1.302	hypothetical protein	PHD	Remodelage de la chromatine
PF11_0429	hypothetical protein	PHD	Remodelage de la chromatine
PF14_0315	hypothetical protein	PHD	Remodelage de la chromatine
PFL0575w	hypothetical protein	PHD	Remodelage de la chromatine
PFL1010c	hypothetical protein	PHD	Remodelage de la chromatine
PFL1905w	hypothetical protein	PHD	Remodelage de la chromatine
PFF1185w	ISWI protein homologue	PHD	Remodelage de la chromatine
PF10_0079	hypothetical protein	PHD	Remodelage de la chromatine
PFB0730w	DNA helicase	SNF2 domain	Remodelage de la chromatine
MAL8P1.65	hypothetical protein	SNF2 domain	Remodelage de la chromatine
PF11_0053	PfSNF2L	SNF2 domain	Remodelage de la chromatine
PFF0225w	DNA helicase	SNF2 domain	Remodelage de la chromatine
PF08_0048	ATP-dependant helicase	SNF2 domain	Remodelage de la chromatine
PF10_0232	hypothetical protein	SNF2 domain	Remodelage de la chromatine
PF13_0308	DNA helicase	SNF2 domain	Remodelage de la chromatine

ARTICLES ORIGINAUX

ARTICLE 1

Transcriptome of 3D7 and its gametocyte-less derivative F12 *Plasmodium falciparum* clones during erythrocytic development using a gene-specific microarray assigned to gene regulation, cell cycle and transcription factors.

Mathieu Gissot, Philippe Refour, Sylvie Briquet, Charlotte Boschet, Stéphane Coupé, Dominique Mazier & Catherine Vaquero.

Accepté à la publication dans *Gene*.

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gene 341 (2004) 267–277

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Transcriptome of 3D7 and its gametocyte-less derivative F12 *Plasmodium falciparum* clones during erythrocytic development using a gene-specific microarray assigned to gene regulation, cell cycle and transcription factors

Mathieu Gissot, Philippe Refour, Sylvie Briquet, Charlotte Boschet, Stéphane Coupé[1], Dominique Mazier, Catherine Vaquero*

INSERM U511, CHU Pitié-Salpêtrière, Université Paris 6, 91 boulevard de l'Hôpital, 75013 Paris, France

Received 28 March 2004; received in revised form 21 June 2004; accepted 5 July 2004
Available online 13 September 2004

Abstract

During the complex life cycle of *Plasmodium falciparum*, through mosquito and human, the erythrocytic cycle is responsible for malarial disease and transmission. The regulation of events that occur during parasite development, such as proliferation and differentiation, implies a fine control of transcriptional activities that in turn governs the expression profiles of sets of genes. Pathways that underline gametocyte commitment are yet poorly understood even though kinases and transcription factors have been assumed to play a crucial role in this event. In order to understand the molecular mechanisms controlling the variation of gene expression profiles that might participate in early gametocytogenesis, the transcriptome of two clones, 3D7 and its gametocyte-less derivative F12, was compared at five time points of the erythrocytic asexual development. We have used a thematic DNA microarray containing 150 PCR fragments, representative of *P. falciparum* genes involved in signal transduction, cell cycle and transcriptional regulation. We identified several genes eliciting different expression profiles among which some implicated in gene regulation or encoding putative transcription factors. The differential expression of transcription factor and kinase transcripts observed in the two clones may enlighten genes that might have a role in impairment of the early gametocytogenesis of the F12 clone.

Keywords: *Plasmodium*; Transcription factor; Microarray; Erythrocytic development; Gametocytogenesis

1. Introduction

Plasmodium is responsible for 1.5–2.7 million deaths annually, mostly children and pregnant women. Among the four species of parasites infecting humans, *Plasmodium falciparum* causes the highest morbidity and mortality. Global efforts to eradicate malaria have failed and no effective vaccine is available. Therefore, insight into the parasite biology and mechanisms of gene and cell cycle regulation during the development of *Plasmodium* is urgently required to provide new targets for the control of malaria. The cell cycle of *Plasmodium* between vertebrate host and mosquito passes by a succession of different stages characterized by an active cellular division (as the exo-erythrocytic and erythrocytic schizogony) and an arrest of proliferation associated to differentiation (as for gametocytes and sporozoites). Actually, two major events of the *P. falciparum* life cycle occur in the erythrocyte, an asexual multiplication responsible for the clinical manifestations and the initiation of sexual differentiation responsible for dissemination of the disease via the mosquito. Those two

Abbreviations: aa, amino acid; CO, Constans; ES, early schizont; ET, early trophozoite; LS, late schizont; LT, late trophozoite; NAP, nucleosome assembly protein; NLS, nuclear localization site; ORF, open reading frame; PCR, polymerase chain reaction; qPCR, real time quantitative PCR; R, ring; RT, reverse transcription; TF, transcription factor; dNTP, deoxyribonucleoside triphosphate; DNase, deoxyribonuclease; RNase, ribonuclease; UTR, untranslated region(s); cDNA, DNA complementary to RNA; IFN, interferon; SDS, sodium dodecyl sulfate; SSC, 0.15 M NaCl/0.015 M Na_3 citrate pH 7.6.

* Corresponding author. Tel.: +33 1 40778111; fax: +33 1 45838858.
E-mail address: vaquero@ext.jussieu.fr (C. Vaquero).
[1] Present address: Laboratoire de Parasitologie-Mycologie, Centre Hospitalier Universitaire Lariboisière-Saint-Louis, 75010 Paris, France.

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.07.004

events are under the control of coordinated expression of distinct sets of genes governed by transcriptional regulation even though post-transcriptional events could also be implicated.

Pathways that control asexual cycle regulation and switch to sexual commitment are yet poorly understood although a number of genes described to command cell cycle in other eukaryotes, e.g. kinases and transcription factors, have been predicted in our laboratory or by the sequencing project of *P. falciparum* (Gardner et al., 2002). Two publications have described the whole genome examination of the transcriptome of HB3 (Bozdech et al., 2003) and 3D7 (Le Roch et al., 2003) clones of *P. falciparum* giving the first global idea of the expression profiles of almost all genes during erythrocytic asexual cycle and at gametocyte stage, therefore opening new perspectives for the discovery of gene implication in those pathways. Various genes were reported to be associated with induction of gamecytogenesis, as *pfg27* and *pfs16* (Dyer and Day, 2000b). Moreover, implication of cyclic AMP-dependent pathway (Inselburg, 1983) and trimeric G proteins (Dyer and Day, 2000a) suggests participation of kinases and transcription factors in this phenomenon. Sexual commitment have been reported to be sensitive to many environmental conditions including host immunity, anti-malarial drugs, host hormones, erythrocyte intracellular environment and autocrine growth control (reviewed in Dyer and Day, 2000b). The parasite response to environment that influences the developmental decision, point out the key role of signaling pathways comprising cell cycle regulators, kinases and transcription factors leading to transcriptional control of sexual commitment. Clones of *P. falciparum* lacking or with a reduced ability to produce gametocytes have been reported. Although those alterations in sexual development have been considered to be irreversible, drug-driven induction of gametocytogenesis was observed in clones thought to be impaired in sexual development (Ono and Nakabayashi, 1990) suggesting a reduced sensitivity to environmental changes. Subtelomeric deletion of chromosome 9 also induces almost total loss of gametocyte production (Day et al., 1993). However, total or partial failure of the ability to produce gametocytes has been shown to occur in clones, derived from 3D7, carrying an intact chromosome 9 and no visible changes in caryotype (Alano et al., 1995). F12 clone has been reported to be one of the gametocyte-less 3D7 derivative. In addition, loss and decrease of mRNA expression of the two earliest landmarks of gametocyte differentiation, the cytoplasmic protein *Pfg27* and the membrane protein *Pfs16*, respectively, have been stated (Alano et al., 1995). However, the role of *Pfg27* in sexual differentiation remains unclear. Its expression have been described to be developmentally regulated (Alano et al., 1996), required for early stage of gametocytogenesis and reported to play a role in a part of *Pfs16* expression induction (Lobo et al., 1994, 1999). Moreover, crystal structure of *Pfg27* raised the hypothesis of potential binding to other proteins such as kinases through SH3 domains and to RNA (Sharma et al., 2003). Nature of the impairment of *pfg27* transcript expression could be due to absence or altered-expression of genes implicated in the control of its expression, since major rearrangement upstream to the *pfg27* coding sequence was not observed in F12 clone (Alano et al., 1996). The absence of expression of *pfg27* transcript and therefore loss of gametocyte production ability remain unclear in F12 clone; whether it has lost the sensitivity to external stimuli or the ability to respond to these stimuli also remains to be elucidated.

In the investigation described here, we constructed a gene-specific microarray encompassing 153 *P. falciparum* PCR products representing genes that might be implicated in regulation of cell cycle and transcription. We hybridized cDNA, obtained from total RNA prepared from five time points of asexual erythrocytic culture of the two clones, 3D7 and F12, and labeled with two distinct radioisotopes. We compared expression profiles of the 153 genes in these two clones grown under normal culture condition. Indeed, we tried to identify genes with differential expression profiles in the F12 clone that does not produce gametocytes and in 3D7 that bear the potentiality to produce gametocytes. This analysis was carried out to underline differentially expressed genes particularly focusing on putative transcription factors. Differential expression of transcription factors and kinases might have a role in the impairment in early stage of gametocytogenesis.

2. Materials and methods

2.1. P. falciparum clones and cultures

P. falciparum 3D7 and F12 (Alano et al., 1995) were provided by Dr. D. Walliker and Dr. P. Alano, respectively, and were cultured as described in Trager and Jensen (1976) with slight modifications. To obtain synchronized erythrocytic stages, ring cultures were enriched by three successive treatments of 5% sorbitol at 44-h intervals. Five stages of development, rings (10–14 h), early (20–22 h) and late (26–28 h) trophozoites, and early (32–34 h) and late (40–44 h post-invasion) schizonts, were determined according to time and morphological analysis by GIEMSA staining. At each indicated stage, parasite culture at 8% parasitemia was collected by centrifugation and used for preparation of transcripts for further analyses.

2.2. Amplification of PCR products and microarray construction

Microarrays were prepared by printing 170 PCR products 300–1500 bp long produced from sequences cloned in the Topo 2.1 vector (Invitrogen), representing mostly genes implicated in signal transduction and cell cycle of *P. falciparum*. PCR fragments from all clones (Eurogentec)

were purified by isopropanol precipitation and resuspended in spotting buffer (Tris 10 mM pH 7.4, EDTA 1 mM, DMSO 50%) adjusted to an average concentration of 200 ng/μl and produced with specific pairs of primers found in the vector; T7 (5′ TAATACGACTCACTATAGGG 3′) and M13R (5′ GGATAACAATTTCACACAGG 3′).

Seventeen spots, representing positive and negative controls such as sequences unrelated to *P. falciparum*, were printed. Alien PCR products from Stratagene, with no apparent cross-similarity with *P. falciparum* genomic sequence, were used for normalization of the reverse transcription. 3D7 genomic DNA was also printed in order to normalize the total amount of each cDNA present during the hybridization. All targets were spotted in duplicate with a Qarray arrayer (Genetix) onto poly-L-lysine-coated slides prepared as described previously (Schena et al., 1995) with a spot spacing of 400 μm, center to center onto a 1 cm^2 area. DNA was cross-linked by UV irradiation at 3 kJ with Stratalinker model 1800 UV Illuminator (Stratagene). Neutralization of free poly-L-lysine was performed by chemical treatment with succinic anhydride (Sigma). Prior to hybridization, the slides were incubated for 1 min at 95 °C to denature the spotted DNA and dehydrated with a 96% ethanol immersion for 1 min and dried by centrifugation.

2.3. Probe labeling, hybridization and data analysis

Parasitized red blood cells were rapidly lysed in TRIZOL (Invitrogen) and total RNA was extracted at different times during erythrocytic development from 3D7 and F12 *P. falciparum* cultures. Concentration was determined by spectrophotometer and integrity of the RNA preparations was verified by ethidium bromide staining on agarose gel and by Agilent 2100 bioanalyser.

Ten micrograms of total RNA extracted at different times of erythrocytic development were labeled during reverse transcription reaction in a sample mixture containing 2 μg of 15-mer poly-T primer (Invitrogen), 5 mM of $MgCl_2$, 10 mM of DTT, 80 U of RNAout (Invitrogen), 0.2 mM of each dCTP, dGTP and dTTP, 50 μCi/μl of ^{35}S-dATP (Amersham, catalog number SJ1134) or 40 μCi/μl of ^{3}H dATP (Amersham, catalog number TRK633). Reaction was performed at 42 °C for 2 h with 250 U of Superscript II reverse transcriptase and subjected to ribonuclease (RNase)-H treatment with 2 U of enzyme for 20 min at 37 °C. Labeled cDNA were purified on QIAquick nucleotide removal columns as indicated by the manufacturer (Qiagen) and eluted in 40 μl. Amount of radioactivity contained in 1 μl of purified eluate was evaluated using a Beckman Coulter LS 6500 apparatus.

For hybridization, the labeled probe mixture corresponding to 1.6×10^6 disintegrations per minute for each cDNA was used per slide. Each labeled cDNA was added to the hybridization buffer (3.5× SSC, 0.3% SDS, 0.5 μg/μl DNA Salmon Sperm and 0.5 μg/μl yeast tRNA), heated to 95 °C for 2 min, cooled down at room temperature and added on the microarray slide under a cover slip (Easy seal, Hybaid). Hybridization was performed in a cassette chamber (Tele-Chem) submerged in a water bath at 62 °C for 14–16 h. Arrays were washed at room temperature in a 2× SSC, 0.1% SDS solution prior to washing in a 2× SSC and in a 0.2× SSC solution successively, each step for 2 min, and finally dried by centrifugation.

Acquisition of arrays was performed as described previously on a Microimager (Biospace Mesure) (Salin et al., 2002), and stopped after 24 h acquisition. Data were collected as an image file and the ARRAYVISION software (Imaging Research) was used for quantification of the hybridization intensities and for normalization. The local background was subtracted from each spot intensity. The intensity of each spot was normalized according to the median value of intensities of spots corresponding to Alien cDNA PCR products on each array in order to assess normalization of the reverse transcription. Furthermore, genomic DNA was used to normalize the amount of cDNA present in the hybridization. For all genes, expression profile was expressed as the $\log_2$ ratio of labeling intensity of experiment over reference. Two independent experiments and one technical replicate were performed.

2.4. Genomic DNA labeling

Genomic DNA (30 ng) isolated from F12 and 3D7 clones were ^{35}S-labeled using a random-primed DNA labeling kit (Amersham). Labeled genomic DNAs were purified using QIAquick columns (Qiagen) and hybridized as described above.

2.5. Real time quantitative RT-PCR (qPCR)

Primers were designed using PrimerExpress software (Applied Biosystem). Total RNA (10 μg) was treated with 10 U of RNase-free deoxyribonuclease (DNase) I (Qiagen), and cDNA synthesis was performed for 50 min at 42 °C using 1 μg of RNA, 50 U of MMLV reverse transcriptase (Superscript II, Invitrogen), 0.5 ng of random hexamers, 5 mM of $MgCl_2$, 20 U of RNaseout (Invitrogen), 10 mM of DTT and 0.25 mM of each dNTP. Finally, reaction was subjected to RNase-H treatment with 2 U of enzyme for 20 min at 37 °C. Real time quantitative PCR reaction was performed in a 20 μl reaction volume containing 5 μl of a dilution of cDNA preparation, 10 μl of SYBR green PCR master mix (Sigma) and 0.375 μM of gene-specific primers (Invitrogen). Amplification and detection of specific products were performed with the MX4000 light cycler (Stratagene) with the following cycle profile: one cycle at 95 °C for 2 min, 40 cycles with 30 s denaturation at 95 °C and 1 min annealing-elongation at 60 °C. Size of each PCR product and number of bands was verified on a 10% acrylamide gel. Two additional reactions were performed, either without reverse transcriptase or without the RNA sample, to verify the absence of DNA contamination. The

quantity of cDNA for each experimental gene was normalized to the 18S concentration (chr5.rRNA-1–18s-A) in each sample. For each time point of each gene, experiments were performed twice and in triplicates. Relative gene expression was expressed via the $\log_2$ of the ratio of the expression time point over ring stage using the $2_T^{-\Delta\Delta C}$ method (Livak and Schmittgen, 2001).

3. Results

3.1. P. falciparum gene-specific microarray, experimental design and validation of the results

We constructed a microarray encompassing 170 PCR products among which 153 amplified from the *P. falciparum* genomic sequences. This thematic microarray could evolve by adding new genes of interest when needed. All sequences to be amplified were localized within the open reading frame (ORF) close to the 3′ end and were sequenced prior to be arrayed on glass slides.

The *Plasmodium* genes were selected for their homology to known proteins previously described in eukaryotes to be involved in genetic regulation from DNA to proteins. Indeed, glass slides were printed with PCR fragments coming from genes encoding putative kinases controlling signal transduction and cell cycle, proteins involved in replication (DNA polymerases, DNA topoisomerases and helicases), proteins involved in general transcription machinery (TATA binding protein, small nuclear ribonucleoprotein, RNA polymerases), transcription factors (TFs) (such as three members of the Myb family, two members of the HMG family and several proteins encompassing zinc finger domains), as well as proteins involved in the stability and the structure of mRNA (ribosomal RNA methylases, helicases), in translation (initiation and elongation factors), and in chromatin remodeling (histone deacetylase, nucleosome assembly protein). Several of these proteins have already been described in the literature, others annotated in databases such as PlasmoDB (www.plasmodb.org) or annotated by our group among which putative transcription factors and kinases. For DNA printing, several negative controls were used among which *pflsa1* only expressed during exo-erythrocytic cycle, unrelated genes as human gamma IFN and HIV-1 gp120 envelope protein, empty plasmid vector used to clone each PCR product and mouse or human genomic DNA. Several *Plasmodium* genes reported to be expressed during asexual erythrocytic cycle, among which *hsp70* (PF08_0054), *H2A* (MAL6P1.249) and *H3* (MAL6P1.248), *kahrp* (PFB0100c), *pcna1* (PF13_0328) and *gbp130* (PF10_0159), were spotted as positive controls at different positions on the microarray. Finally, PCR products corresponding to the Alien (Stratagene) synthetic RNAs, added into total RNA samples to normalize the effectiveness of the cDNA probe production, as well as two different *Plasmodium* genomic DNA concentrations, used to normalize the total amount of each cDNA present in the hybridization, were also printed.

We performed radioactive labeling of the cDNA reverse transcribed from all the total RNA samples, in order to minimize background signal, enhance signal detection of low abundant mRNA species and avoid signal saturation. This experimental approach was also further adapted to very low concentrations of biological material (Refour et al., submitted for publication). A CCD camera determined, in a real time manner, the signals obtained after hybridization of the microarray with the two differentially ^{35}S or ^{3}H labeled cDNA. For each *P. falciparum* clone, two independent technical replicates as well as two different biological RNA preparations from independent culture were used for cDNA synthesis. A representative picture of co-hybridization data obtained with the cDNA labeled with the two radioisotopes is presented in Fig. 1. Negative controls showed no signal as for *pflsa1* (PF10_0356) (a), human genomic DNA (b), human gamma IFN (c) and HIV-1 gp120 protein (d), in contrast to positive controls, genes expressed during asexual erythrocytic cycle (1–3). In addition, three PCR products along the sequence of PF14_0316 ORF were spotted to control the PCR product sequence effect on the expression profile (4–6). Moreover, the same PCR product of PF11_0098 ORF (7–8) was spotted twice to control the localization effect and all gave similar signals. Duplicate spots appeared to give similar signal levels. In addition, self-hybridization between ring stage cDNA labeled with ^{35}S and ring stage cDNA labeled with ^{3}H gave ratio very close to one (data not shown).

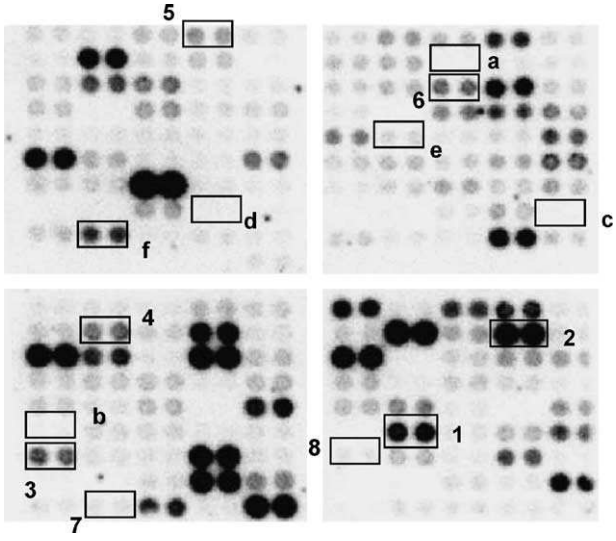


Fig. 1. A representative picture of a co-hybridization of two cDNA labeled in the presence of ^{35}S dATP and ^{3}H dATP. Framed duplicated spots represent (a) *pflsa1* (PF10_0356), (b) human genomic DNA (20 ng/μl), (c) human gamma IFN, (d) HIV gp120 protein, (e) Alien 1 from Stratagene, (f) Alien 2 from Stratagene; (1) *pfhsp70* (PF08_0054), (2) *histone H2A* (MAL6P1.249), (3) *gbp130* (PF10_0159); (4, 5, 6) three different PCR products for PF11_0098; (7, 8) the same PCR fragment for PF14_0316 spotted twice.

When performing the first time course experiment for the 3D7 clone, we selected five genes as positive controls for which expression has been previously described. We compared throughout development the peaks of expression obtained either by our microarray or qPCR to published data (Table 1). The different expression profiles were equivalent and in good agreement with the published data, therefore validating the quality of our experimental approach.

Finally, prior to the first microarray differential experiments, genomic DNA from F12 clone were labeled and hybridized to the microarray in order to verify that the level of homology between F12 and 3D7 sequences was sufficient to determine and compare the abundance of all messengers of each clone. Actually, all immobilized PCR products amplified from 3D7 clones hybridized to labeled genomic DNA extracted from F12 (data not shown). In addition, we verified by real time quantitative RT-PCR that total RNA prepared from the F12 different stages was devoid of *pfg27* transcript (data not shown).

3.2. Comparison between 3D7 and F12 expression profiles

The expression profiles of 3D7 clone and its gametocyteless derivative F12 clone were monitored during the erythrocytic cycle. Highly synchronized cultures were harvested at five time points during the asexual cycle. Five stages of development, rings (10–14 h), early (20–22 h) and late (26–28 h) trophozoites, and early (32–34 h) and late (40–44 h post-invasion) schizonts, were assessed according to time and morphological analysis. The cDNA obtained from ring total RNA of each *Plasmodium* clone was selected as the reference RNA, labeled with ^{35}S dATP and co-hybridized with the cDNA labeled with ^{3}H dATP prepared from total RNA extracted at each indicated time of development. The five time points are rings (R), early (ET) and late trophozoites (LT), and early (ES) and late schizonts (LS). For all genes, the steady state RNA profile was expressed as the $\log_2$ ratio of labeling intensity of experiment (R, ET, LT, ES, or LS) over reference (R). In addition, signal intensity was normalized twice, first with the Alien cDNA to normalize the RT efficacy and second with the genomic *Plasmodium* DNA in order to normalize the amount of labeled cDNA loaded during the hybridization process.

Among the 153 *Plasmodium* genes spotted on the array, 23 genes including the seven negative controls, reported to be expressed only during exo-erythrocytic cycle, gave no detectable signal. Among the remaining 130 genes, 16 exhibited too low expression intensity and therefore no reproducible results for the three replicates assayed. When the 114 expression profiles obtained for 3D7 clone were compared to the two recently published transcriptome data for HB3 (Bozdech et al., 2003) and 3D7 clones (Le Roch et al., 2003), more than 80% (82,5%) of our expression results (see Supplementary data 1) were similar to at least one of these published data.

We compared the expression profiles of the 114 genes expressed during the erythrocytic cycle of the two clones and found that 106 genes shared similar profiles, consistent with the fact that F12 clone derived from 3D7 (Alano et al., 1995) (see Supplementary data 2). However, eight genes displayed differential expression and those results were confirmed by real time quantitative PCR experiments. These genes were clustered in three groups highlighting the differences between the two clones (Fig. 2). First, two kinases genes (*MAL6P1.146* and *PF13_0258*) with nearly no modification in expression throughout 3D7 cycle in contrast to what was observed during F12 development with a maximal expression at LT for *MAL6P1.146* and ES for *PF13_0258* (Fig. 2A). Second, two genes (*PFC0485w* and *PFB0865w*) presented opposite profiles. The *PFC0485w* gene, a putative Ser/Thr kinase, was found to have a maximal expression during ring stage and a minimal expression during schizont stage in 3D7 clone in contrast to the expression profiles of F12 clone. The expression of *PFB0865w*, a putative small nuclear ribonucleoprotein, peaked in 3D7 clone in LT and was always higher than the level observed in rings, whereas mRNA relative amount decreased from ring to schizont in F12 (Fig. 2B).

Third, the maximal peak of expression of genes represented in Fig. 2C was reached for 3D7 at different stages of development forwarded or delayed when compared to the F12 clone. *PFI0930c*, a nucleosome assembly protein and *PFC0805w*, a putative DNA-directed RNA polymerase II largest subunit, showed a maximal peak at ET stage in 3D7 clone whereas the maximal peak was observed later in F12 cycle. On the contrary, in the case of *PF07_0073*, a seryl-tRNA synthetase and *PF14_0383*, a Constans-like putative transcription factor, the maximal

Table 1
Comparison of the stage where maximal expression was observed for five control genes analyzed in 3D7 clone by microarray experiments, real time quantitative PCR or published data

Gene	Technical procedure		
	Microarray[a]	qPCR[b]	References
Histone H2A MAL6P1.249	S[c]	S	S (Lobo and Kumar, 1999)
KAHRP PFB0100c	R	R	R (Lanzer et al., 1992a)
GBP130 PF10_0159	LT	LT	LT (Lanzer et al., 1992b)
Pfs 40[d] PF11_0098	S	S	S (La Greca et al., 1997)
PCNA1 PF13_0328	ES	ES	ES (Horrocks et al., 1996)

[a] The stage of maximal expression was determined from the microarray data for the 3D7 clone.
[b] By real time quantitative RT-PCR experiments.
[c] S stands for schizonts, R for rings, LT for late trophozoite and ES for early schizonts.
[d] Results are given for the three different PCR products as stated in Materials and methods.

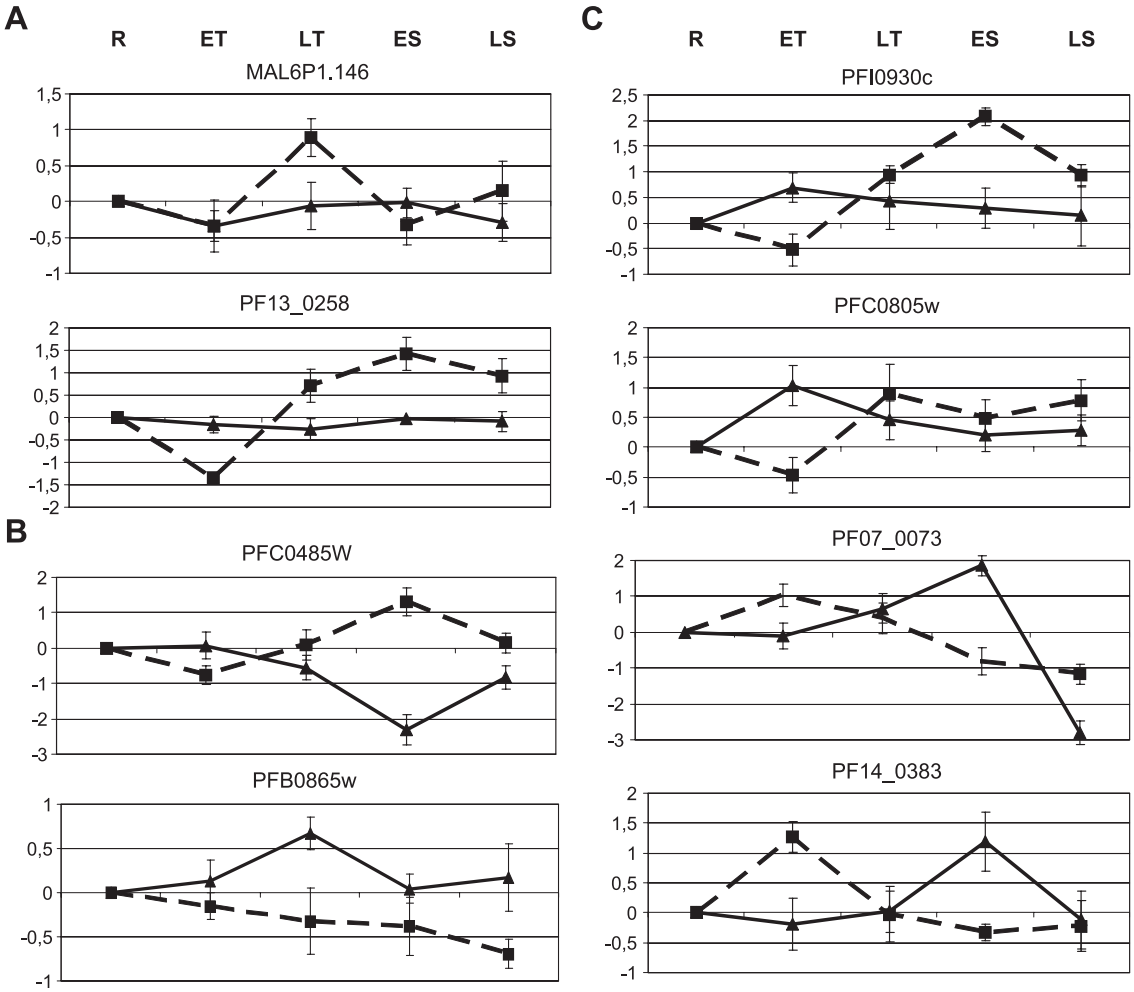


Fig. 2. Graphic representation of the expression profile of eight genes showing differential expression between the two clones 3D7 and F12. For each gene, the steady state RNA profile was expressed as $\log_2$ ratio of labeling intensity of experiment (any time point) over reference (ring). 3D7 expression profile is represented by a continuous line with black triangles and F12 expression profile by a dash line with black squares. R stands for rings, ET for early trophozoites, LT for late trophozoites ES for early schizonts and LS for late schizonts, stated at the top of the figure. For each clone, the average of three experiments are shown as well as standard deviation. The genes have been clustered in three groups according to their expression profiles: (A) two kinases MAL6P1.146 and PF13_0258, (B) PFB0865w, a putative small nuclear ribonucleoprotein, and PFC0485w, a putative serine/threonine kinase, and, (C) PFC0805w, a putative DNA-directed RNA polymerase II largest subunit, PFI0930c, a nucleosome assembly protein, PF14_0383, a Constans-like putative transcription factor, and PF07_0073, a seryl-tRNA synthetase.

peak was reached in the ES in 3D7 and in ET for the F12 clone (Fig. 2C).

3.3. Comparison of the transcriptional machinery proteins in 3D7 and F12 clones

We particularly focused our attention on the putative transcription factors and proteins of the transcriptional machinery, some previously annotated by our group, among which several are under molecular and functional investigation.

All 15 transcription factors were developmentally regulated during erythrocytic asexual cycle of the two clones with essentially similar patterns of expression. In Fig. 3, we selected five putative TFs presenting different expression patterns throughout erythrocytic development however similar in 3D7 and F12 clones by microarray experiment (Fig. 3, panel A) and confirmed the data by real time quantitative PCR experiment (Fig. 3, panel B). *Pfhmg2* (PFL0290w) was expressed at a similar level throughout the asexual cycle with slight variations when compared to ring stage. In contrast, the relative amount of *pfphD2* declined strongly more than four times from R to ES re-increasing thereafter. For the last three transcripts, maximal expression was observed in ET for *pfmyb3* (PF10_0143), in ES for *pfphDB* (PFL1905w) and LS for *pfkrox* (MAL13P1.76). All the messengers coding for these TF were developmentally regulated peaking at different times during the erythrocytic asexual cycle.

However, as already stated in Fig. 2, several messengers encoding various proteins of the transcription machinery exhibited differential expression profiles when comparing the 3D7 to F12 clone. Those transcripts encode a putative nuclear ribonucleoprotein (PFB0865w), a DNA-directed

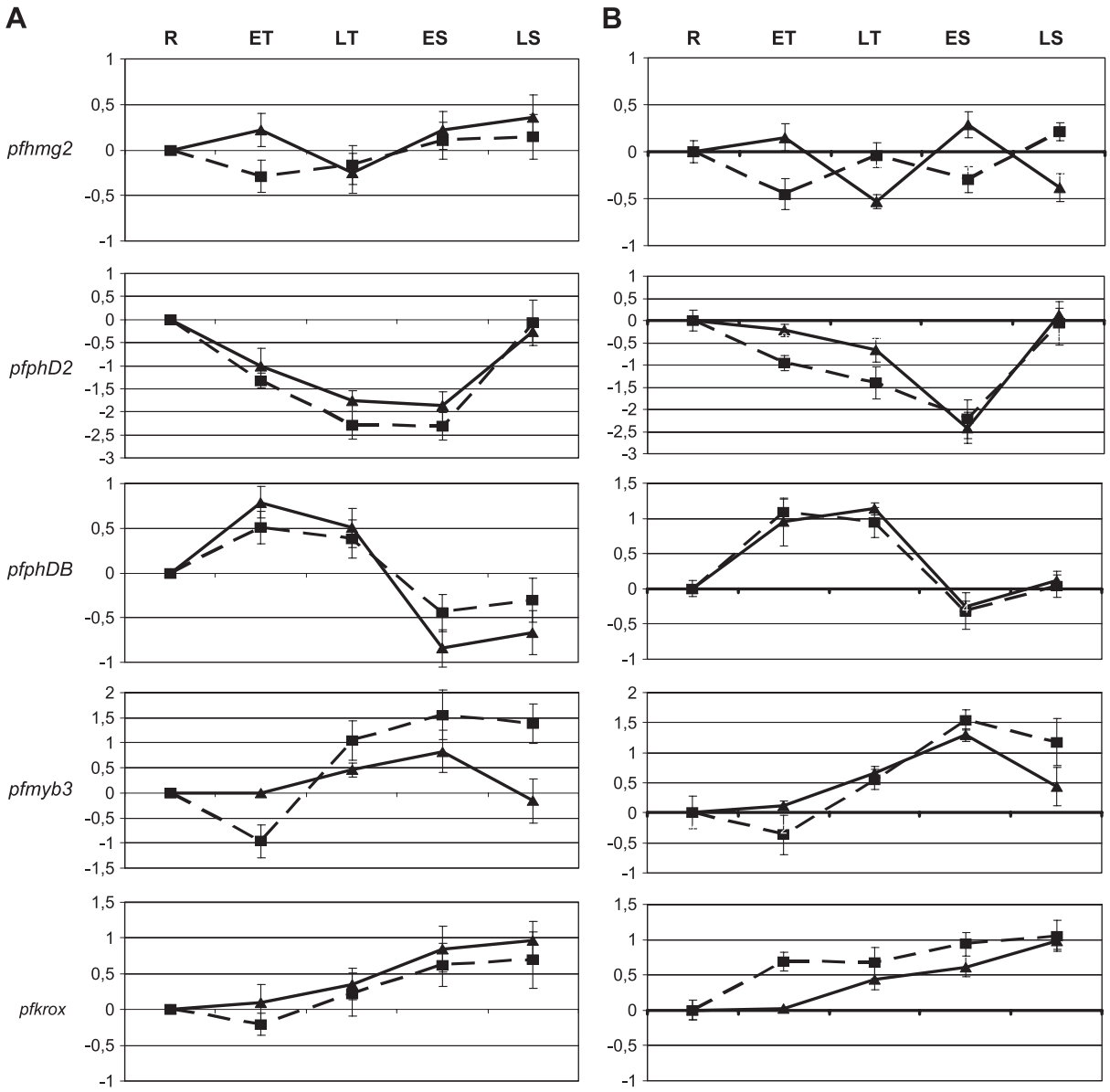


Fig. 3. Graphic representation of the expression profile of five putative transcription factors in the two clones 3D7 and F12 monitored via microarray and qPCR. For each gene, expression profile was expressed and represented as mentioned in Fig 2. (A) Expression profile obtained from microarray experiments. (B) Expression profile derived from real time quantitative PCR experiments.

RNA polymerase II (PFC0805w), a nucleosome assembly protein (NAP) (PFI0930c) and a Constans-like transcription factor (PF14_0383). These messengers and proteins should be analysed to determine if their differential expression could participate in the transcriptional regulation of varied sets of genes during the asexual parasite cycle.

4. Discussion

In order to determine differential transcript expression during erythrocytic development of *P. falciparum* clones 3D7 and its gametocyte-less derivative F12, we have used a gene-specific microarray printed with 153 PCR products representing genes potentially implicated in cell cycle and transcriptional regulation. F12 clone has been isolated as a gametocyte non-producer from a long-term culture of the 3D7 clone. In contrast to 3D7 clone, F12 is impaired for early stages of gametocytogenesis without apparent difference in chromosomes number or size (Alano et al., 1995). Due to the genomic proximity of these two clones, one might expect that differential gene expression during the erythrocytic asexual cycle would participate in the early stage of gametocytogenesis including sensitivity to environment stimuli known to induce differentiation into gametocyte. Therefore, we designed a gene-specific microarray to study and compare temporal expression of roughly 150 messengers, during erythrocytic asexual development, of

highly synchronized cultures to identify some genes that might play a role in the parasite differentiation.

Before performing the time course transcriptome experiments we verified the clone properties, the quality of synchronization procedure and the feasibility of using F12 derived cDNA with a microarray encompassing 3D7 derived PCR product. First, we monitored expression of *pfg27* transcript by real time quantitative RT-PCR and verified that no expression of the transcript could be detected in F12 in contrast to 3D7. Second, we carefully synchronized the parasite cultures, by three successive treatments with sorbitol at 44-h intervals. Subsequently, the five stages of erythrocytic asexual development were harvested with a minimum of 90% enrichment at each stage. Finally, we verified that all the PCR fragments obtained after amplification of the 3D7 genomic sequence and printed on the array could cross-hybridize with the genetic material prepared from F12 clone.

The radioactive labeled cDNAs were prepared from total RNA extracted at the indicated times of asexual life cycle. Formerly, the quality of the RNAs was controlled by Agilent analyzer and by fractionation on agarose gel. We performed radioactive-labeling of the cDNA probes, in order to minimize the background signal, avoid signal saturation and enhance signal detection especially for transcription factor mRNA that are believed to have a low abundance level. At the level of the microarray experimental protocol, we controlled the novel radioactive labeling and hybridization procedures carried out with the two probes (Salin et al., 2002; Refour et al., submitted for publication). Indeed, this was necessary to certify the reliability and reproducibility of this novel experimental approach.

Fig. 1 presents a representative picture corresponding to a co-hybridization with two cDNA labeled with ^{35}S dATP and ^{3}H dATP. Printing and hybridization were verified by the (i) appropriate responses given by the positive and negative controls, (ii) quality of the duplicates, and (iii) similarity of the signals given by the same PCR product (PF11_0098) spotted twice at different places in the array as well as that given by three different PCR products along the PF14_0316 ORF. In addition, each raw signal was normalized twice, first with the Alien cDNAs and second with the genomic *Plasmodium* DNA. Finally, some of the microarray data were verified using qPCR and compared with the results reported in the literature (Table 1).

The comparison of the steady state profiles of mRNA expression between the two clones, 3D7 and F12, led us to list 106 genes with essentially similar profiles and only eight genes with differential expression (Fig. 2). Additionally, we focused our attention on all the transcripts probably implicated in the transcriptional machinery and in particular those coding for transcription factors. Indeed, among the transcripts of the 15 putative transcription factors, most of them gave similar profiles throughout erythrocytic development of 3D7 and F12 clones, as shown for five of them verified by quantitative real time PCR (Fig. 3A and B). This is not surprising since F12 was derived from 3D7. In addition, the conserved transcriptional profiles of several of these putative transcription regulators during erythrocytic cycle in 3D7, F12 and HB3 clones (Bozdech et al., 2003), favour their significant role for asexual growth and not for the impairment of sexual differentiation observed in the F12 clone.

All genes and in particular those encoding kinases and proteins of the transcriptional machinery displaying different expression profiles, were compared to results reported in the Scripps/GNF Malaria Array data (Le Roch et al., 2003) and the De Risi *P. falciparum* HB3 time course microarray data (Bozdech et al., 2003). Most of our data obtained with the 3D7 clone were in accordance with these two published experiments. However, some differences were observed that might be due to the clones features or/and to the cDNA preparation and labeling protocol (fluorescence and use of RNA amplification).

Even though most of the transcript profiles were similar between 3D7 and F12 clones, differential profiles were observed for genes potentially coding for kinases. First, two kinases, *MAL6P1.146* (haem-regulated inhibitory (HRI) kinase) and *PF13_0258* (a putative serine/threonine protein kinase) were clustered since expression was quite similar throughout 3D7 cycle in contrast to a maximal peak of expression occurring in the F12 clone, at LT for *MAL6P1.146* and ES for *PF13_0258* (Fig. 2A). Since *PF13_025* expression peaked in the schizont stage for 3D7 (Scripps/GNF Malaria Array data) and HB3 clones as well as for F12 clone, this gene does not appear to be essential for either asexual or sexual growth.

MAL6P1.146, also known as PfPK4, has been annotated as a protozoan eIF-2α-related protein kinase due to its homology to HRI kinase, and reported to participate to the inhibition of protein synthesis, via the phosphorylation of the small subunit of the eIF-2 initiation factor (Clemens, 2001). Indeed, activity of the recombinant PfPK4 has been described to be inhibited by haemin (Mohrle et al., 1997). In erythrocytes infected by the K1 isolate, the level of the MAL6P1.146 protein appeared quite similar during the parasite cycle even though its cellular distribution varied (Mohrle et al., 1997). Since PfPK4 accumulates in rhoptries, it probably plays a role in invasion and early development of the ring stage. In addition, erythrocytes contain small concentrations of free haemin and this level is increased in aged erythrocytes and pathological cells of sickle cell anaemia (Shaklai et al., 1985). The inhibition of PfPK4 activity by haemin might play a role in sensing internal erythrocyte environment. Since differential expression takes place between 3D7 and F12 clones and since enhanced gametocyte production occurs in young erythrocytes and in blood from patients with sickle cell anaemia (Trager and Gill, 1992), PfPK4 might be implicated in commitment of the young ring to gametocytogenesis by sensing internal erythrocyte environment. Over-expression of this gene in trophozoite might be at least in part

responsible for a decreased sensitivity to erythrocyte internal environment leading to its inability to respond to the commitment stimulus.

Second, *PFC0485w* encoding a putative Ser/Thr kinase, displayed opposite expression profiles (Fig. 2B) with, a maximal expression for F12 in early schizonts, and a minimal for expression 3D7 clone. Interestingly, expression has also been shown to be increased in 3D7 gametocytes (Le Roch et al., 2003), and therefore, might be involved in gametocytogenesis providing an explanation for the gametocyte-less phenotype of F12 clone.

Finally, *PF07_0073*, a seryl-tRNA ligase, expression has been shown to be maximal in the early schizonts in 3D7 as also observed in HB3 in contrast to F12 where expression was maximal at early trophozoite stage (Fig. 2C) as seen for 3D7 (Scripps/GNF Malaria Array data, sorbitol synchronization). The expression profile of this gene may not be crucial for sexual development since it is different in the gametocyte producing clones.

In addition to kinases, several transcripts coding for proteins involved in transcriptional machinery showed differential profiles between the two clones even though most of the TF transcripts investigated presented similar expression during erythrocytic development.

The relative level of *PFB0865w* transcript, encoding a putative small nuclear ribonucleoprotein probably involved in RNA splicing, was higher than ring in 3D7 clone, in good agreement with the Scripps/GNF Malaria Array data, and continuously lower than ring in F12, as stated for the HB3 clone. Since HB3 clone is a gametocyte producer, this gene may not play a crucial role in controlling gametocyte production.

The maximal peak of expression of *PFC0805w* (a putative DNA-directed RNA polymerase II largest subunit) was reached at different stages of development, delayed in F12 when compared to 3D7. The delay observed in F12 clone may have very limited effects considering the crucial role of RNA polymerase II in transcription.

Finally, two putative transcriptional regulators *PFI0930c* and *PF14_0383*, showed distinct expression profiles in the two clones. Expression of *PFI0930c* transcript encoding a protein also known as PfB7, bearing a NAP motif, is not greatly altered throughout 3D7 parasite while it is markedly enhanced in F12 early schizonts. Proteins bearing a NAP motif have been shown to act as histone chaperones, shuttling both core and linker histones from their site of synthesis in the cytoplasm to the nucleus (Akey and Luger, 2003). These proteins may be involved in regulating gene expression by interaction either with chromatin remodelling factors therefore promoting transcriptional silencing (Singer et al., 1998) or with architectural factors such as HMGB proteins interacting with nucleosomes and promoting their sliding (Agresti and Bianchi, 2003). Interestingly, as we have annotated various HMGB proteins and since they are expressed during erythrocytic cycle (to be published), PfB7 might be responsible for the modulation of the functional activity of these proteins. We have identified orthologs protein of *PFI0930c* in the four *Plasmodium* species accessible in PlasmoDB that showed a very high level of similarity (about 95% of the amino acid content), as stated before (Birago et al., 1996), enlightening the crucial role of this protein (data not shown). Moreover, evidence of two differently-sized mRNAs, encoding this protein, have been obtained in *P. falciparum* (Pace et al., 1998). In *P. berghei*, two different lengths of the 5′ untranslated region (UTR) have been proposed to explain these two forms of mRNAs, one accumulating during gametocyte differentiation and the other specific for asexual growth. Since the two transcripts share an identical coding sequence, the PCR product used in this study detects either asexual or sexual specific transcript. Different species of *PbB7* mRNA have been hypothesized to be the results of two different promoters specific for asexual or sexual growth. Therefore, *PFI0930c* regulation has been demonstrated to be linked to gametocyte differentiation in *P. berghei* (Pace et al., 1998) and one can assume that it is also the case in *P. falciparum*. Differential expression in 3D7 and F12 clone may therefore be a consequence of differential expression of transcription factors and kinases implicated in the gametocyte differentiation. Indeed, chromatin remodelling may play a role in the impairment of F12 to produce gametocytes.

In the case of *PF14_0383* transcript coding for a Constans-like transcription factor, the maximal peak was obtained in the early schizont stage in 3D7 when it was observed during early trophozoite stage in F12 clone (Fig. 2C). The PF14_0383 protein comprises two B-box motifs as seen in Constans (CO) proteins (Robson et al., 2001) and two nuclear localization signals (NLS) in tandem. Indeed, the B-box zinc finger motifs are present in a large number of proteins involved in transcription regulation, usually associated with RING finger and coil–coil motifs to form a tripartite motif, and if mutated, associated with human disease and cancer (Torok and Etkin, 2001). In plants, CO belongs to a family of transcription factors characterized by two zinc-finger B-boxes in tandem in their N-termini, believed to mediate protein–protein interaction (Robson et al., 2001; Samach et al., 2000). CO is involved in transcriptional regulation of flowering in *Arabidopsis thaliana* and promotes the expression of specific sets of genes (Samach et al., 2000). Again, we identified protein orthologs in the four *Plasmodium* species accessible in PlasmoDB (data not shown) showing conserved region located in the two B-boxes and in two other locations of the coding region including the two putative NLS. The first B-box matches the exact consensus (C *x*2 C *x*8 C *x*7 C *x*2 C *x*4 H *x*8 H) derived from CO B-box (Robson et al., 2001) where *x* represents any amino acid (aa). In addition, two other amino acids conserved in CO motif were shown, if mutated, to result in a flowering phenotype mutant of *A. thaliana* (Koornneef et al., 1991), and were conserved in the PF14_0383 first B-box motif. The second B-box appeared

quite similar to the canonical B-box consensus [C *x*2 H x(4–9) C *x*2 C *x*4 C *x*2 H (C)] (Torok and Etkin, 2001). Bioinformatics analyses of the two B-boxes appeared to reveal a genuine CO-like transcription factors. Although its function as a transcription factor, has not yet been demonstrated, it possesses highly conserved motifs that have been involved in transcription regulation. Differential expression of this gene between the two clones may promote differential expression of sets of genes implicated in gametocytogenesis.

This study is the first attempt to analyse, by using a gene-specific microarray, the expression profile of messengers encoding proteins implicated in the cell cycle and transcription machinery, among which potential transcription factors not involved in the basal transcription but in the modulation of transcription level. Our results led us to identify, during the erythrocytic development of the 3D7 and F12 clones, transcription factors and kinases differentially expressed. They might in turn lead to differential expression of different sets of genes directly or indirectly implicated in the gametocyte differentiation. Nonetheless, true functional involvement in gametocytogenesis remains to be demonstrated. More work is needed and already engaged in order to decipher the potential involvement of each differentially expressed gene in F12 clone gametocyteless phenotype.

Acknowledgements

We are indebted to Dr. Walliker and Dr. Alano for the kind gift of the *P. falciparum* clones. We are grateful for the microarray facilities provided by L. Galio and A. Doukani at P3S (Pitié-Salpêtrière). M.G., P.R. and C.B. were financially supported by the Ministère de l'Education Nationale, de la Recherche et de la Technologie and the PAL+ program. This work was supported by INSERM, France, to C.V. and D.M.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2004.07.004.

References

Agresti, A., Bianchi, M.E., 2003. HMGB proteins and gene expression. Curr. Opin. Genet. Dev. 13, 170–178.

Akey, C.W., Luger, K., 2003. Histone chaperones and nucleosome assembly. Curr. Opin. Struck. Biol. 13, 6–14.

Alano, P., Roca, L., Smith, D., Read, D., Carter, R., Day, K., 1995. *Plasmodium falciparum*: parasites defective in early stages of gametocytogenesis. Exp. Parasitol. 81, 227–235.

Alano, P., Silvestrini, F., Roca, L., 1996. Structure and polymorphism of the upstream region of the pfg27/25 gene, transcriptionally regulated in gametocytogenesis of *Plasmodium falciparum*. Mol. Biochem. Parasitol. 79, 207–217.

Birago, C., Pace, T., Barca, S., Picci, L., Ponzi, M., 1996. A chromatin-associated protein is encoded in a genomic region highly conserved in the *Plasmodium* genus. Mol. Biochem. Parasitol. 80, 193–202.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., De Risi, J.L., 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol. 1, 5.

Clemens, M.J., 2001. Initiation factor eIF2 alpha phosphorylation in stress responses and apoptosis. Prog. Mol. Subcell. Biol. 27, 57–89.

Day, K.P., Karamalis, F., Thompson, J., Barnes, D.A., Peterson, C., Brown, H., Brown, G.V., Kemp, D.J., 1993. Genes necessary for expression of a virulence determinant and for transmission of *Plasmodium falciparum* are located on a 0.3-megabase region of chromosome 9. Proc. Natl. Acad. Sci. U. S. A. 90, 8292–8296.

Dyer, M., Day, K., 2000a. Expression of *Plasmodium falciparum* trimeric G proteins and their involvement in switching to sexual development. Mol. Biochem. Parasitol. 110, 437–448.

Dyer, M., Day, K.P., 2000b. Commitment to gametocytogenesis in *Plasmodium falciparum*. Parasitol. Today 16, 102–107.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, 498–511.

Horrocks, P., Jackson, M., Cheesman, S., White, J.H., Kilbey, B.J., 1996. Stage specific expression of proliferating cell nuclear antigen and DNA polymerase delta from *Plasmodium falciparum*. Mol. Biochem. Parasitol. 79, 177–182.

Inselburg, J., 1983. Stage-specific inhibitory effect of cyclic AMP on asexual maturation and gametocyte formation of *Plasmodium falciparum*. J. Parasitol. 69, 592–597.

Koornneef, M., Hanhart, C.J., van der Veen, J.H., 1991. A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. Mol. Gen. Genet. 229, 57–66.

La Greca, N., Hibbs, A.R., Riffkin, C., Foley, M., Tilley, L., 1997. Identification of an endoplasmic reticulum-resident calcium-binding protein with multiple EF-hand motifs in asexual stages of *Plasmodium falciparum*. Mol. Biochem. Parasitol. 89, 283–293.

Lanzer, M., de Bruin, D., Ravetch, J.V., 1992a. A sequence element associated with the *Plasmodium falciparum* KAHRP gene is the site of developmentally regulated protein–DNA interactions. Nucleic Acids Res. 20, 3051–3056.

Lanzer, M., de Bruin, D., Ravetch, J.V., 1992b. Transcription mapping of a 100 kb locus of *Plasmodium falciparum* identifies an intergenic region in which transcription terminates and reinitiates. EMBO J. 11, 1949–1955.

Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., Winzeler, E.A., 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science.

Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25, 402–408.

Lobo, C.A., Kumar, N., 1999. Differential transcription of histone genes in asexual and sexual stages of *Plasmodium falciparum*. Int. J. Parasitol. 29, 1447–1449.

Lobo, C.A., Konings, R.N., Kumar, N., 1994. Expression of early gametocyte-stage antigens Pfg27 and Pfs16 in synchronized game-

tocytes and non-gametocyte producing clones of *Plasmodium falciparum*. Mol. Biochem. Parasitol. 68, 151–154.

Lobo, C.A., Fujioka, H., Aikawa, M., Kumar, N., 1999. Disruption of the Pfg27 locus by homologous recombination leads to loss of the sexual phenotype in *P. falciparum*. Mol. Cell. 3, 793–798.

Mohrle, J.J., Zhao, Y., Wernli, B., Franklin, R.M., Kappes, B., 1997. Molecular cloning, characterization and localization of PfPK4, an eIF-2alpha kinase-related enzyme from the malarial parasite *Plasmodium falciparum*. Biochem. J. 328 (Pt. 2), 677–687.

Ono, T., Nakabayashi, T., 1990. Gametocytogenesis induction by ammonium compounds in cultured *Plasmodium falciparum*. Int. J. Parasitol. 20, 615–618.

Pace, T., Birago, C., Janse, C.J., Picci, L., Ponzi, M., 1998. Developmental regulation of a *Plasmodium* gene involves the generation of stage-specific 5′ untranslated sequences. Mol. Biochem. Parasitol. 97, 45–53.

Refour, P., Siau, A., Gissot, M., Briquet, S., Boschet, C., Mazier, D., Vaquero, C., 2004. High-accuracy of nanogram total RNA amounts for gene profiling without amplification (submitted for publication).

Robson, F., Costa, M.M., Hepworth, S.R., Vizir, I., Pineiro, M., Reeves, P.H., Putterill, J., Coupland, G., 2001. Functional importance of conserved domains in the flowering-time gene CONSTANS demonstrated by analysis of mutant alleles and transgenic plants. Plant J. 28, 619–631.

Salin, H., Vujasinovic, T., Mazurie, A., Maitrejean, S., Menini, C., Mallet, J., Dumas, S., 2002. A novel sensitive microarray approach for differential screening using probes labelled with two different radioelements. Nucleic Acids Res. 30, e17.

Samach, A., Onouchi, H., Gold, S.E., Ditta, G.S., Schwarz-Sommer, Z., Yanofsky, M.F., Coupland, G., 2000. Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. Science 288, 1613–1616.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

Shaklai, N., Shviro, Y., Rabizadeh, E., Kirschner-Zilber, I., 1985. Accumulation and drainage of hemin in the red cell membrane. Biochim. Biophys. Acta 821, 355–366.

Sharma, A., Sharma, I., Kogkasuriyachai, D., Kumar, N., 2003. Structure of a gametocyte protein essential for sexual development in *Plasmodium falciparum*. Nat. Struct. Biol. 10, 197–203.

Singer, M.S., Kahana, A., Wolf, A.J., Meisinger, L.L., Peterson, S.E., Goggin, C., Mahowald, M., Gottschling, D.E., 1998. Identification of high-copy disruptors of telomeric silencing in *Saccharomyces cerevisiae*. Genetics 150, 613–632.

Torok, M., Etkin, L.D., 2001. Two B or not two B? Overview of the rapidly expanding B-box family of proteins. Differentiation 67, 63–71.

Trager, W., Gill, G.S., 1992. Enhanced gametocyte formation in young erythrocytes by *Plasmodium falciparum* in vitro. J. Protozool. 39, 429–432.

Trager, W., Jensen, J.B., 1976. Human malaria parasites in continuous culture. Science 193, 673–675.

ARTICLE 2

High-Mobility-Group box nuclear factors of *Plasmodium falciparum*.

Sylvie Briquet, Charlotte Boschet, Mathieu Gissot, Emilie Tissandié, Elisa Sevilla, Jean-François Franetich, Isabelle Thierry, Zuhal Hamid, Catherine Bourgoin & Catherine Vaquero.

Accepté à la publication dans *Eukaryotic Cell*.

EUKARYOTIC CELL, Apr. 2006, p. 672–682 Vol. 5, No. 4
1535-9778/06/$08.00+0 doi:10.1128/EC.5.4.672–682.2006

High-Mobility-Group Box Nuclear Factors of *Plasmodium falciparum*†

Sylvie Briquet,[1]* Charlotte Boschet,[1] Mathieu Gissot,[1]‡ Emilie Tissandié,[1] Elisa Sevilla,[1] Jean-François Franetich,[1] Isabelle Thiery,[2] Zuhal Hamid,[1]§ Catherine Bourgouin,[2] and Catherine Vaquero[1]*

INSERM, U511, Université Pierre et Marie Curie, Paris VI, Centre Hospitalo-Universitaire de la Pitié-Salpêtrière, Paris, France,[1] and Biologie et Génétique du Paludisme, CEPIA (Centre de Production et d'Infection des Anophèles), Institut Pasteur, Paris, France[2]

Received 30 November 2005/Accepted 31 January 2006

In eukaryotes, the high-mobility-group (HMG) nuclear factors are highly conserved throughout evolution and are divided into three families, including HGMB, characterized by an HMG box domain. Some HMGB factors are DNA structure specific and preferentially interact with distorted DNA sequences, trigger DNA bending, and hence facilitate the binding of nucleoprotein complexes that in turn activate or repress transcription. In *Plasmodium falciparum*, two HMGB factors were predicted: PfHMGB1 and PfHMGB2. They are small proteins, under 100 amino acids long, encompassing a characteristic HMG box domain closely related to box B of metazoan factors, which comprises two HMG box domains, A and B, in tandem. Computational analyses supported the conclusion that the *Plasmodium* proteins were genuine architectural HMGB factors, and in vitro analyses performed with both recombinant proteins established that they were able to interact with distorted DNA structures and bend linear DNA with different affinities. These proteins were detected in both asexual- and gametocyte-stage cells in Western blotting experiments and mainly in the parasite nuclei. PfHMGB1 is preferentially expressed in asexual erythrocytic stages and PfHMGB2 in gametocytes, in good correlation with transcript levels of expression. Finally, immunofluorescence studies revealed differential subcellular localizations: both factors were observed in the nucleus of asexual- and sexual-stage cells, and PfHMGB2 was also detected in the cytoplasm of gametocytes. In conclusion, in light of differences in their levels of expression, subcellular localizations, and capacities for binding and bending DNA, these factors are likely to play nonredundant roles in transcriptional regulation of *Plasmodium* development in erythrocytes.

Malaria is the most important parasitic disease in the world, and of the 300 to 500 million cases each year, approximately 2 million people die. Among the four species of malaria parasites infecting humans, *Plasmodium falciparum* causes the highest morbidity and mortality. Global efforts to eradicate malaria have failed, and there is presently no effective vaccine available. Therefore, a greater understanding of parasite biology throughout development is urgently needed in order for novel therapeutic strategies to control malaria to be proposed.

During the erythrocytic life cycle, intense multiplication of parasites takes place, as well as gametocyte differentiation associated with cell cycle arrest. These different developmental pathways require the coordinated and modulated expression of diverse sets of genes, involving transcriptional, epigenetic, and posttranscriptional regulation. Currently, it is commonly accepted that general mechanisms involved in gene regulation in eukaryotes also operate in *P. falciparum* (25, 32, 33). Nevertheless, elucidation of the molecular mechanisms involved in transcriptional regulation in *Plasmodium* is still challenging. Even if very little is known about the *cis*- and *trans*-regulatory elements of the parasite, *Plasmodium* genes exhibit the bipartite structure of eukaryotic promoters, i.e., a basal promoter regulated by upstream regulatory elements (25) that present some homology with the binding sites of eukaryotic transcription factors (TF). The recent completion of the genome sequence of *P. falciparum* revealed a high proportion of orphan proteins (60% of the open reading frames [ORFs] have no match with any of the annotated sequences listed in the data banks [18]). These data might contribute to the low numbers of recognizable, orthologous TF (11). However, it is reasonable to assume that in *Plasmodium* the interplay between regulatory elements and TF, whose availability (49) presumably modulated throughout parasite development, governs also the level of RNA synthesis.

In eukaryotes, in addition to general TF also annotated in *Plasmodium* (10, 23, 37, 38, 43, 44), the factors involved in transcriptional regulation can be divided into factors interacting either with specific DNA sequences (42) or with DNA structures. The latter include the nonhistone proteins of the high-mobility-group (HMG) superfamily (7, 58, 62), which is divided into three families of proteins in line with their characteristic functional motifs (8): HMGA, which interacts with the AT hook; HMGN, which interacts with the nucleosomes; and HMGB, which encompasses one or several copies of the HMG box DNA binding domain (for a review, see reference 7). HMG proteins are present in all metazoan phyla, plants, and yeast and have also been reported in unicellular parasites, including trypanosomes (15, 45), schistosomes (21), and *Plasmodium* (29). They are quite abundant proteins, one molecule

* Corresponding author. Mailing address: INSERM, U511, Université Pierre et Marie Curie, Paris VI, Centre Hospitalo-Universitaire de la Pitié-Salpêtrière, 91 boulevard de l'Hôpital, 75013 Paris, France. Phone: 33 (0) 1 40 77 81 14. Fax: 33 (0) 1 45 83 88 58. E-mail for Sylvie Briquet: briquet@ext.jussieu.fr. E-mail for Catherine Vaquero: vaquero@ext.jussieu.fr.

† Supplemental material for this article may be found at http://ec.asm.org/.

‡ Present address: Albert Einstein College of Medicine, New York, N.Y.

§ Present address: INMO, University of Gezira, Sudan.

for 10 to 15 nucleosomes in vertebrates. It is assumed that the wrapping of DNA by histones and nonhistone proteins, including the HMG proteins, controls the access of the TF to their target sites on nucleosomes (31).

HMGB factors are highly conserved throughout evolution, and their HMG box domain is composed of around 80 amino acids (aa) folded in three α-helices arranged in an L shape (3, 66). In vertebrates, the HMGB proteins generally present two boxes, A and B, and also basic N- and C-terminal extensions and a rather long C-terminal acidic tail (58). Despite their low sequence homology, both boxes (A and B) present a well-conserved L-shaped structure, even though their DNA binding and bending capacities may display some differences (28, 69). In lower eukaryotes, either the basic extension (*Drosophila melanogaster*) or the negatively charged tail (*Saccharomyces cerevisiae*) is missing, in contrast to plant HMGs (60), which possess both extensions, albeit of different lengths. The basic domains appear to play a role in the stabilization of HMGB-mediated DNA bending. In contrast, the role of the acidic tail remains elusive and may be shaped to interact with the positive charges of histones (for a review, see reference 62). Two subfamilies of HMGB, with either DNA sequence specificity (SOX, SRY, TCF, MATA) or structure specificity (HMGB per se), have been identified. The latter preferentially interacts with distorted DNA sequences and triggers DNA bending, hence altering the positioning of nucleosomes on the DNA fiber, thereby controlling the level of transcription (31). Finally, a linker histone H1 (27), via its interaction with the DNA linker between two nucleosomes, increases the compactness of the chromatin (53), impairing interactions between DNA and TF and therefore repressing gene transcription (24). In contrast, the HMGB proteins appear to be associated with active chromatin (48), increasing nucleosome sliding and target site accessibility and thereby enhancing transcription.

Lately, these proteins, historically known as nuclear proteins, have been reported to be released from mammalian cells and to act as mediators of the immune response and as potent macrophage-activating factors (for reviews, see references 14, 41, and 46).

In *Plasmodium*, very few TF have been annotated (PlasmoDB and our group [19]) and characterized (6, 20). In *Plasmodium falciparum*, four potential HMG factors have been annotated, including one previously reported for the FCQ27 (29) and FCC1/HN parasite clones. Two of these factors, PfHMGB1 and PfHMGB2, were investigated during the erythrocytic cycle to evaluate their molecular implications in transcription regulation.

MATERIALS AND METHODS

***Plasmodium falciparum* culture.** The 3D7 clone of *P. falciparum* was provided by D. Walliker and was grown in human erythrocytes, as described by Trager and Jensen (61), except that the culture medium contained 0.5% Albumax instead of human serum. Gametocytes were produced from the gametocyte-producing isolate NF54 (provided by W. Eling) as aforementioned, with 10% AB human serum, using the automated culturing system developed by Ponnudurai et al. (52). Parasitemia was monitored on Giemsa-stained blood smears. Thirteen to 15 days after induction of gametocytogenesis, gametocytes were monitored for maturity by observation under a microscope (1 × 100 immersion) of microgamete exflagellation, with the addition of 5 μl of gametocyte culture in a 10-μl drop of human serum.

Antibodies. Primary antibodies used to characterize the PfHMGB1 and PfH MGB2 proteins were obtained after immunization of BALB/c mice with either recombinant protein (50 μg, twice at 2-week intervals). Sera with high levels of anti-PfHMGB1 and anti-PfHMGB2 activity were collected after the first booster inoculation (day 30) and selected for immunochemical analyses.

Annotation. Before the sequence of the *P. falciparum* genome was completed in 2002, PfHMGB1 was annotated within chromosome 12 by homology to a consensus of the HMG box domains of around 50 disparate eukaryotic sequences. Then, the HMG box domain of PfHMGB1, used as a query against the *Plasmodium* database, allowed the annotation of PfHMGB2 and PfHMGB3 within chromosomes 8 and 12, respectively, while that of PfHMGB2 allowed the annotation of PfHMGB4 within chromosome 13. The presence of relevant HMG box domains in these proteins was checked with MotifScan (16).

Molecular cloning. The Pf*hmgb1* and Pf*hmgb2* ORF sequences were amplified from genomic DNA of the *P. falciparum* 3D7 clone, with forward (5′-GGTGG ATCCATGAAGAATACAGGAAAAGAAG-3′ and 5′-GGTCCCGGGGCCA ATTTAAGCTTTCATTTTC-3′) and backward (5′-ATTGGATCCATGGCTTC AAAATCTCAAAA-3′ and 5′-ATTGGTACCTTATTCTTGATTTTTCTTTC-3′) primers, respectively, using PCR conditions with an elongation temperature of 60°C, as described previously (57). Fragments of 294 bp and 300 bp, corresponding to the complete Pf*hmgb1* and Pf*hmgb2* ORF sequences, were cloned directly into the pGEM-T Easy vector (Promega) and pCR II-Topo (Invitrogen), respectively, and then sequenced with the ABI Prism kit (Perkin Elmer).

Northern blot analysis. Total RNA was purified from isolated parasites with TRIzol (Invitrogen), and the integrity of the RNA preparation was monitored by ethidium bromide staining on an agarose gel and analysis with an Agilent bioanalyzer. The Pf*hmgb1* and Pf*hmgb2* transcripts were characterized from 20 μg of 3D7 total RNA by Northern blotting, according to the Ultrahyb protocol of Ambion, with a 65°C hybridization temperature. Antisense α-^{32}P riboprobes were prepared as previously described (50) from the pGEMT-*pfhmgb1* and pCRII Topo-*pfhmgb2* vectors by using T7 and SP6 RNA polymerase (Promega), respectively.

Expression and purification of recombinant proteins. The pQE30 vector (QIAGEN) was used to express the recombinant proteins (rePfHMGB1 and rePfHMGB2) in *Escherichia coli* as His_6-tagged proteins with BamHI-XmaI- and BamHI-KpnI-digested inserts from the above-mentioned vectors, respectively. The bacterial strain SG 13009, harboring the Pf*hmgb1* or Pf*hmgb2* expression construct, was grown at 37°C in 200 ml of 2YT medium containing 100 μg/ml ampicillin as well as 50 μg/ml kanamycin. Expression was induced with 0.1 mM IPTG (isopropyl-β-D-thiogalactopyranoside) for 3 h at 37°C, and collected cells were solubilized in sonication buffer S (25 mM Tris-HCl [pH 8], 300 mM NaCl, 10 mM imidazole, 10 mM β-mercaptoethanol, 0.5% Triton X-100), 1 ml per mg of dry pellet, in the presence of lysozyme and a 1/25 final dilution of a protease inhibitor cocktail tablet (Roche). Purification of His-PfHMGB proteins was performed essentially as previously described with Ni-nitrilotriacetic acid agarose beads (QIAGEN) (34). After three washes with buffer S supplemented with 20 mM imidazole, bound proteins were eluted with either 250 mM or 50 mM imidazole in 20 mM Tris-HCl (pH 8)–300 mM NaCl for rePfHMGB1 or rePfHMGB2, respectively.

EMSA with synthetic four-way DNA junctions. The partially complementary oligonucleotides 1 to 4, previously described (5, 67), were used to create four-way DNA junctions (4H), as well as 3H and 2H. In addition, the radiolabeled leg 1 containing cruciform 4H was fractionated and eluted from a 5% polyacrylamide gel in 0.5× Tris-borate-EDTA (TBE) buffer. Electrophoretic mobility shift assays (EMSA) were performed by incubating labeled 4H with increasing amounts (0 to 25 μM) of rePfHMGB1 or rePfHMGB2 in a 10-μl final volume for 20 min at room temperature. This dose-response experiment was carried out to determine the amounts of protein necessary to create a major 4H-PfHMGB complex. In the competition assay, 100- and 500-fold molar excesses of cold complete 4H or incomplete 3H and 2H were added to the reaction for an additional 20 min. Samples were run on a 6.5% polyacrylamide gel in 0.5× TBE buffer at 120 V. The vacuum-dried gels were autoradiographed with intensifying screens at −80°C overnight.

Ligase-mediated circularization assay. The circularization assay was based on existing protocols (9, 67). Linear DNA fragments of 123 bp were γ-^{32}P 5′ end labeled and preincubated with increasing concentrations (0.25 to 100 μM) of rePfHMGB1 or rePfHMGB2 for 20 min at room temperature in 1× DNA ligase buffer (New England Biolabs [NEB]) in a final volume of 10 μl. DNA circularization was generated with T4 DNA ligase (NEB), and linear DNA was subsequently digested by exonuclease III (NEB). Samples were treated with proteinase K (Invitrogen) and run on a 6.5% polyacrylamide gel in 0.5× TBE buffer at 120 V. The gels were vacuum dried and autoradiographed.

Preparation of parasite NE, CE, and total extracts and Western blot analysis. Nuclear extracts (NE), cytoplasmic extracts (CE), and total lysates were prepared from 50 ml of red blood cells at 10 to 12% parasitemia infected with 3D7 asexual-stage cells, as described by Osta et al. (49). Proteins (12 μg) were run on

12% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and subjected to Western blotting experiments (see Fig. 4b) after transfer onto polyvinylidene difluoride membranes (Bio-Rad). For sexual-stage analysis, 2 ml of NF54 culture was harvested on day 13 after induction of gametocytogenesis, with 13.8% parasitemia including 55% mature gametocytes (stage V). After erythrocytic lysis in phosphate-buffered saline–0.15% saponin, gametocytes were collected and resuspended in 40 μl of Laemmli buffer. Half of this sample was fractionated for detection of either protein, PfHMGB1 or PfHMGB2. A similar amount of asexual-stage culture was run in parallel (see Fig. 4c). The blots were probed with a 1:2,000 dilution of anti-PfHMGB1 or anti-PfHMGB2 serum, followed by incubation with a peroxidase-conjugated anti-mouse immunoglobulin G (IgG) antibody (Sigma), and revealed by chemiluminescence (Perkin Elmer) or the Supersignal West Femto kit (Pierce). Negative controls were performed with preimmune sera and positive controls with the recombinant proteins. Finally, an anti-HSP70 serum (55) was used as a positive control for CE (see Fig. 4b) and for normalization of protein loading between asexual/sexual stages (QuantiScan Biosoft 2.1).

Immunofluorescence assay. For localization of PfHMGB1 and PfHMGB2 in *P. falciparum*, asexual and sexual stages obtained as described previously were washed twice in RPMI, fixed in 2% paraformaldehyde for 20 min at room temperature, and laid on poly-L-lysine-coated multiwell glass slides. Endogenous fluorescence was quenched with 75 mM NH_4Cl for 10 min. The parasites were permeabilized with 0.5% Triton X-100 for 3 min, blocked with 0.5% bovine serum albumin in phosphate-buffered saline for 1 h, and then incubated with a 1:200 dilution of primary antibodies (anti-PfHMGB1, anti-PfHMGB2, anti-HSP70) for 1 h at room temperature, followed by incubation with a fluorescein isothiocyanate-conjugated anti-mouse IgG antibody (Sigma) and DAPI (4′,6′-diamidino-2-phenylindole) for 1 h before examination by fluorescence microscopy using a Leica DM RD fluorescent microscope equipped with UV and fluorescein filters. The time of exposure was selected in relation to the intensity of the immunofluorescence obtained with each antibody. Images were acquired with Lucia 4.7 and merged with Adobe Photoshop.

Nucleotide sequence accession numbers. The nucleotide sequences of the genes encoding the following proteins have been assigned the corresponding PlasmoDB accession numbers: PfHMGB1, PFL0145c; PfHMGB2, MAL8P1.72; PfHMGB3, PFL0290w; PfHMGB4, MAL13P1.290.

RESULTS

Several predicted factors in *Plasmodium* belong to the HMGB family. Four putative HMG proteins in *P. falciparum* were predicted by sequence homology, as described in Materials and Methods. The proteins appeared to belong to the HMGB family and were named PfHMGB1 to PfHMGB4. The first two, PfHMGB1 and PfHMGB2, are small proteins, under 100 aa in length, while PfHMGB4 has been predicted to encode a 160-aa-long protein. All three encompass only one HMG box domain. In contrast, PfHMGB3 is a larger protein (2,284 aa), with two HMG box domains and several additional putative functional motifs, including one Myb domain (2). The two small proteins, PfHMGB1 and PfHMGB2, are presented herein. It is worthy of note that in contrast to most eukaryotic HMGBs, displaying N- and C-terminal extensions of diverse lengths (58) and reported to bear functional roles, the two *Plasmodium* factors have only a short basic extension upstream of the HMG box domain and no acidic C-terminal tail.

Soullier et al. (56) revealed by phylogenetic analysis that the HMGB factors can be separated into two clearly defined subgroups: (i) the SOX/SRY/MATA/TCF family, whose members are able to bind specific linear DNA sequences, and (ii) the HMGB/UBF family, whose members interact with high affinity to distorted DNA structures. After the addition of *Plasmodium* and *Babesia bovis* sequences, the same analysis was performed several times with different random number seeds, and we were able to assign PfHMGB1 and PfHMGB2 to the subgroup of HMGB proteins characterized by DNA structure specificity (see supplemental material S1) (17, 59). This finding was strengthened by the alignment shown in Fig. 1, performed with several eukaryotic HMG box domains and two sets of complete HMGB1 and HMGB2 sequences issued from diverse *Plasmodium* species. The HMGB of *Plasmodium* possessed two of the three determinants reported to determine the structural DNA specificity (47), that is, the presence in positions 10 and 32 (according to the residue numbering of *Drosophila* HMG-D) of a serine and a hydrophobic residue, respectively. Therefore, we assigned the *Plasmodium* factors to the architectural HMGB family. Finally, a phylogenetic tree was issued from the HMG box domains of PfHMGB1 and PfHMGB2 and those of various metazoan proteins containing two boxes, A and B, in tandem (see supplemental material S2). This analysis, performed several times with different random number seeds, revealed that the HMG box sequences of both *Plasmodium* proteins are more similar to box B than to box A of proteins with two HMG box domains.

The three-dimensional structure of the two *Plasmodium* factors was modeled by homology using as template the structure of box B of the Chinese hamster HMG1 protein (PDB file 1hsn [54]). Four α-helices, called 1, 1′, 2, and 3 (underlined residues in Fig. 1 corresponding to the three α-helices, I, II, and III, stated at the top of Fig. 1 for *D. melanogaster* HMG-D), were predicted to fold in an L shape (see supplemental material S3) (12, 30) in the PfHMGB1 sequence from His 19 to Tyr 90 and in the PfHMGB2 sequence from Ala 23 to Gln 98.

Therefore, all of these computational analyses agreed in suggesting that the two PfHMGB proteins were genuine factors of the HMGB family and may therefore behave as potential architectural factors.

PfHMGB1 and PfHMGB2 interact with 4H. Several sets of in vitro assays were used to validate the computational identification. After the genes were cloned, expression of rePfHMGB1 and rePfHMGB2 was carried out in *Escherichia coli* and His proteins were purified as described in Materials and Methods.

First, increasing amounts of both recombinant rePfHMGB1 and rePfHMGB2 (up to 25 μM) were incubated with radiolabeled complete 4H to analyze the formation of 4H-rePfHMGB1 and 4H-rePfHMGB2 complexes by EMSA. When 3 μM rePfHMGB1 (Fig. 2a) or 0.6 μM rePfHMGB2 (Fig. 2b) was added to the reaction, the 4H labeled cruciform became incorporated into a major 4H-PfHMGB1 or 4H-PfHMGB2 retarded band, respectively. These amounts of recombinant factors were used for the subsequent EMSA experiments.

Second, the binding specificities of the 4H-PfHMGB1 and 4H-PfHMGB2 complexes were analyzed by competition experiments, that is, by adding a 100- or 500-fold molar excess of either complete (4H) or incomplete (3H and 2H) cold DNA junctions, after DNA-protein complex formation. A 500-fold excess of the integral 4H structure abolished the interaction of either protein, whereas a 100-fold molar excess shifted the 4H-PfHMGB2 interaction completely but the 4H-PfHMGB1 interaction only weakly (Fig. 2c and d). In addition, the competition with cold 3H and 2H was ineffective upon analysis of PfHMGB1, in contrast to the substantial competition observed in the presence of PfHMGB2. Thus, binding of PfHMGB1 to 4H required an intact crossover-containing structure. Alto-

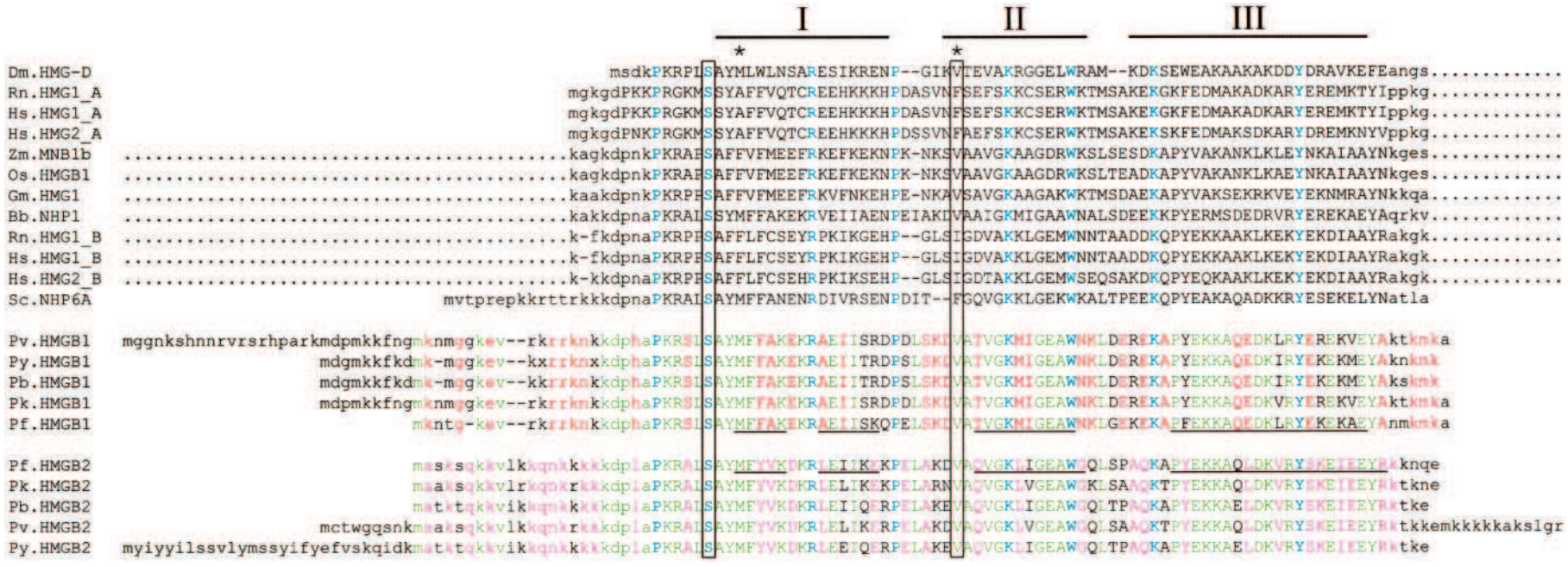


FIG. 1. Multiple alignment of HMG box domains of several eukaryotic HMG proteins and complete sequences of *Saccharomyces* NHP6A and HMGB1 and HMGB2 of diverse *Plasmodium* species. Dashes and dots represent gaps and missing sequences, respectively. Uppercase letters represent HMG box domains identified by the program MotifScan (16). Identities shared by all HMGB sequences presented herein are indicated in bold blue print, by all *Plasmodium* sequences in green, by all *Plasmodium* HMGB1 sequences in red, and by all *Plasmodium* HMGB2 sequences in pink. The boxed amino acids are two crucial determinants that differ between the sequence-specific and the structure-specific HMG box domains: here, a serine and a hydrophobic residue, as found in all non-sequence-specific HMG proteins, whereas all sequence-specific HMG proteins present an asparagine and a hydrophilic residue at these positions. I, II, and III represent the three α-helices of the *D. melanogaster* HMG-D structure (PDB file 1qrv [47]), and asterisks indicate the two residues of *D. melanogaster* HMG-D that intercalate in the DNA minor groove. Underlined residues in PfHMGB1 and PfHMGB2 represent the four α-helices modeled by homology with box B of the Chinese hamster HMG1 protein, used as a template (PDB file 1hsn [54]). Abbreviations are as follows: Bb, *Babesia bovis*; Dm, *Drosophila melanogaster*; Gm, *Glycine max*; Os, *Oryza sativa*; Pb, *Plasmodium berghei*; Pf, *Plasmodium falciparum*; Pk, *Plasmodium knowlesi*; Pv, *Plasmodium vivax*; Py, *Plasmodium yoelii*; Rn, *Rattus norvegicus*; Sc, *Saccharomyces cerevisiae*; Zm, *Zea mays*.

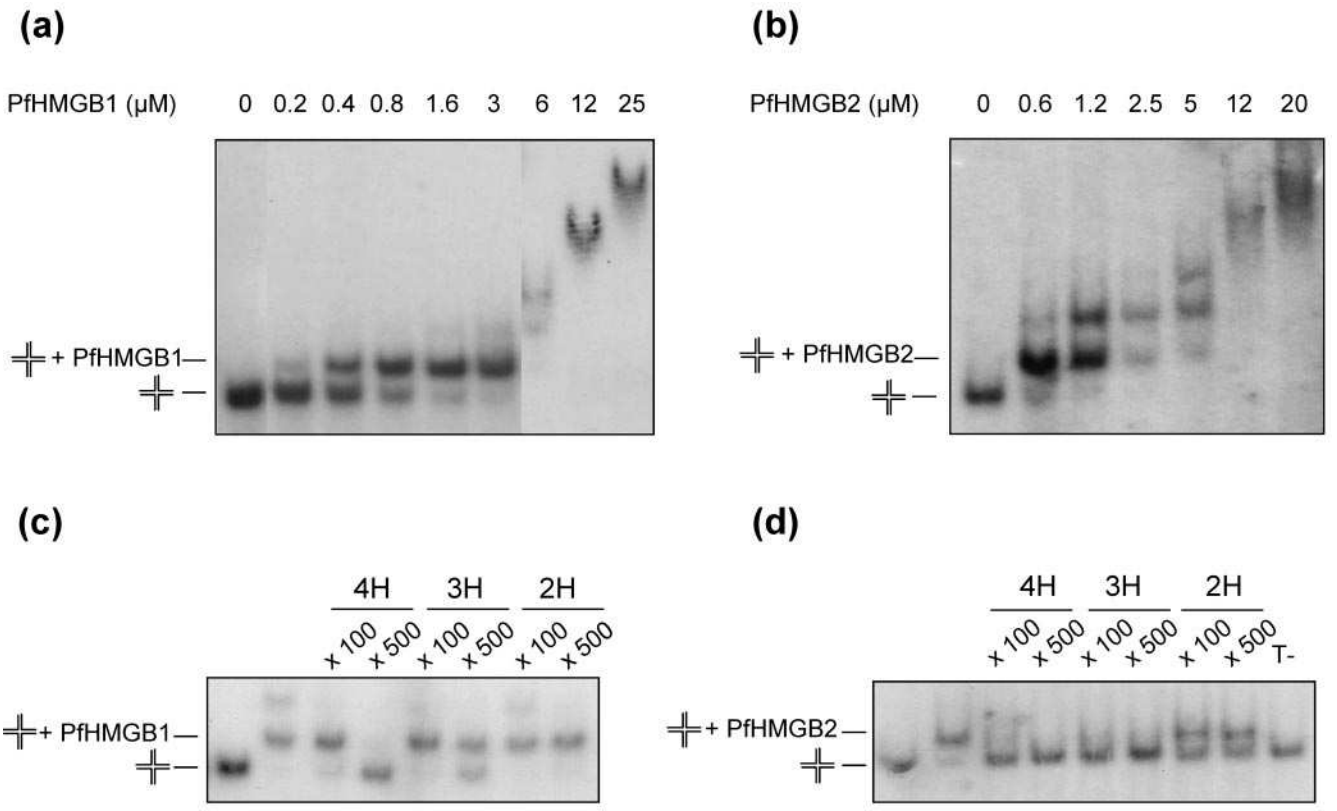


FIG. 2. EMSA interaction between cruciform DNA and either rePfHMGB1 (a) or rePfHMGB2 (b). Increasing concentrations (0 to 25 μM) of rePfHMGB proteins were incubated with radiolabeled 4H. Competition EMSA experiments were performed between the rePfHMGB1-4H (c) or rePfHMGB2-4H (d) complex and various DNA competitors. After incubation of the rePfHMGB proteins with labeled 4H, either cold incomplete junctions (3H and 2H) or complete 4H were added to the reaction at 100- and 500-fold molar excesses, as indicated above each lane. ✣, cruciform DNA. The first lanes correspond to cruciform DNA migration without protein interaction (a to d), and the second lanes correspond to a cruciform DNA-rePfHMGB complex retarded band without competition (c and d).

gether, these results indicated that the two recombinant proteins were able to bind distorted DNA in vitro, even though the nature of their interaction with 4H was quite different, being more efficient and specific for PfHMGB1 than for PfHMGB2.

PfHMGB1 and PfHMGB2 induce DNA bending. We compared the efficiencies of increasing concentrations of rePfH MGB1 and rePfHMGB2 to bend and in turn promote T4 DNA ligase-mediated circularization of a labeled synthetic linear DNA fragment. Indeed, when ligase was added to the labeled fragment of around 125 bp, several bands appeared, including a circular DNA form resistant to exonuclease III (Fig. 3a and b). In the presence of exonuclease III alone, which digests only linear DNA molecules, a marked decrease in all labeled bands was observed, showing that in the absence of PfHMGB proteins, only small amounts of minicircles, if any, were produced. In contrast, in the presence of ligase and increasing amounts of PfHMGB proteins, the quantity of minicircles was quite increased, suggesting that both proteins were capable of enhancing DNA flexibility and hence DNA circularization. The capacity for DNA bending is thus an intrinsic property of PfHMGB1 and PfHMGB2. Nevertheless, rePfHMGB1 once again showed greater efficacy (Fig. 3a), since it started to promote circularization at 0.25 μM, at a concentration 10-fold lower than that of rePfHMGB2 (3 μM), and the maximum signal was reached with 1 μM compared with 50 μM rePfH MGB2. Moreover, the signal observed with 50 to 100 μM PfHMGB2 (Fig. 3b) was far weaker than that observed with 2 μM PfHMGB1 (Fig. 3a).

Pf*hmgb1* and Pf*hmgb2* transcripts are expressed during the *P. falciparum* erythrocytic cycle. In order to determine the presence of Pf*hmgb1* and Pf*hmgb2* RNA and to characterize them molecularly, we performed a Northern blotting analysis of total RNA prepared from infected red blood cells (Fig. 4a). The lengths of Pf*hmgb1* and Pf*hmgb2* mRNA were estimated at 1.3 and 1.1 kb, respectively. The integrity and quality of total RNA extracts of 3D7 were verified after ethidium bromide staining of the gel. As already described for many other *Plasmodium* messengers, the 5′ and 3′ untranslated regions of both transcripts are quite long, as the coding regions comprise fewer than 300 nucleotides.

PfHMGB1 and PfHMGB2 proteins are present in the *P. falciparum* nucleus. Localization of PfHMGB1 and PfHMGB2 was analyzed by Western blotting analysis using CE and NE of *P. falciparum* asexual stages and specific antisera raised against each recombinant protein. The NE and CE prepared from 3D7 parasites (12 μg), as well as the recombinant proteins rePfH MGB1 and rePfHMGB2 (50 ng), were fractionated by SDS-PAGE, and after transfer, the membranes were developed

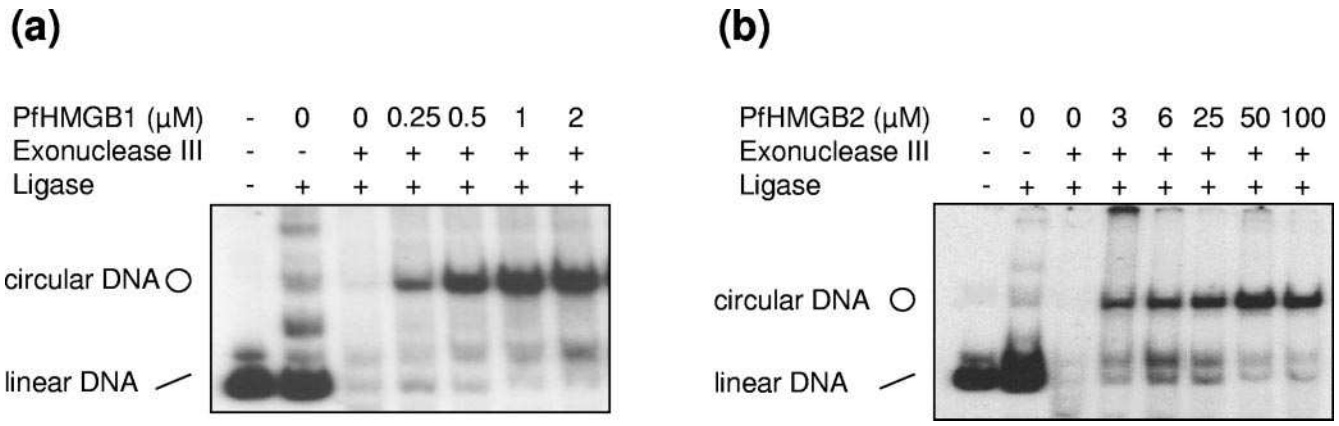


FIG. 3. DNA bending and ligase-mediated circularization assay with either rePfHMGB1 (a) or rePfHMGB2 (b). The γ-^{32}P 5′ end-labeled 123-bp DNA fragment was preincubated with increasing amounts of rePfHMGB1 (0 to 2 μM) or rePfHMGB2 (0 to 100 μM), followed by ligation with T4 DNA ligase. The ligation products were subjected to electrophoresis after exonuclease III treatment. T4 DNA ligase was added to all samples except that loaded in the first lane. All samples were treated with exonuclease III except samples of the first two lanes. The migration positions of 123-bp linear and 123-bp minicircular DNAs are indicated at left.

with the specific antisera (Fig. 4b). The specificities of the two antibodies were verified. No cross-reaction was observed, as His-PfHMGB1 and His-PfHMGB2 were recognized only by their respective antisera (Fig. 4b, lanes 4 to 5 and lanes 8 to 9 for His-PfHMGB1 and His-PfHMGB2, respectively). The control experiments performed with the two preimmune sera gave no signal (data not shown). Both PfHMGB1 and PfHMGB2 were clearly detected in the NE (lanes 3 and 7), whereas same protein loading of CE does not give any detectable signal (lanes 2 and 6). The quality of the CE preparations was controlled by detection of the HSP protein (70 kDa) (lane l). The apparent molecular mass of both PfHMGB proteins in the nuclear extracts, around 12 kDa, was in good agreement with the theoretical molecular masses of 11.3 kDa and 11.5 kDa for

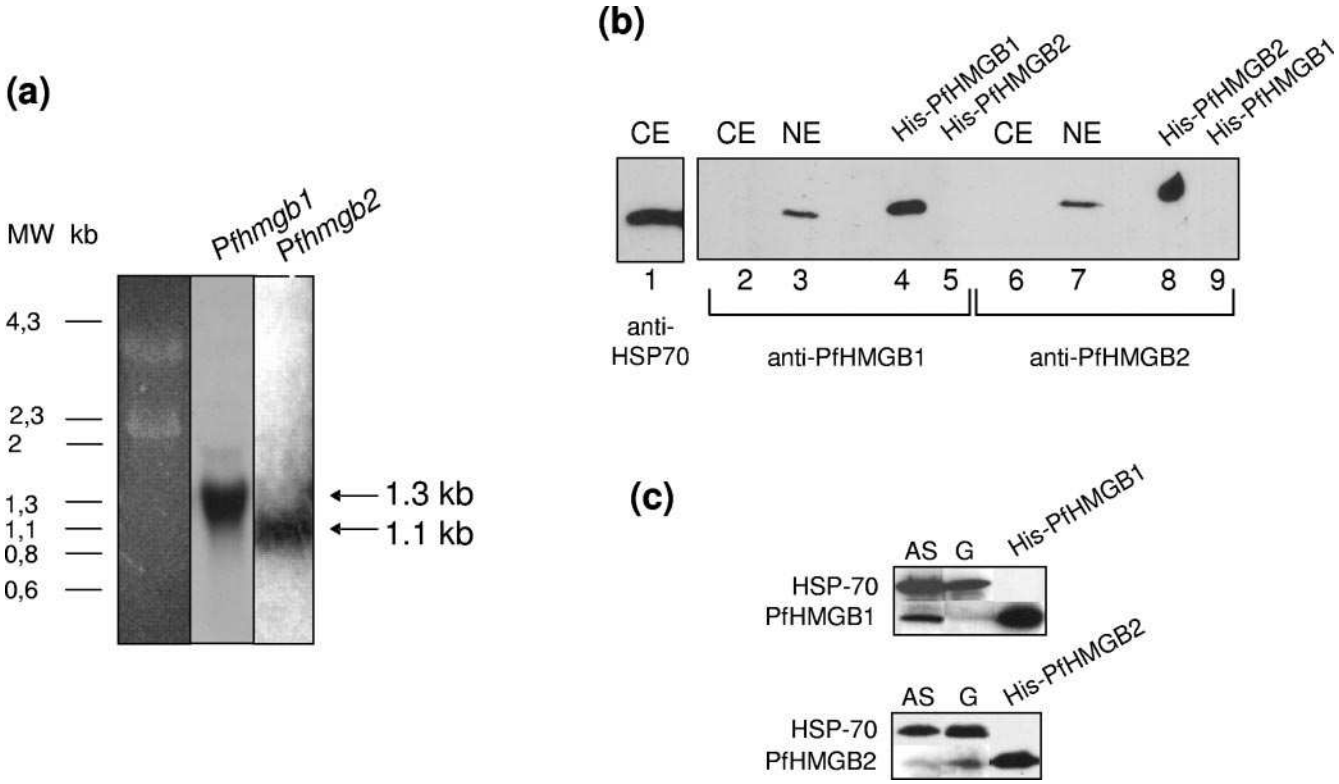


FIG. 4. Characterization of P*fhmgb1* and P*fhmgb2* transcripts by Northern blotting (a) and of their corresponding proteins, His-PfHMGB1 and His-PfHMGB2, by Western blotting (b). By SDS-PAGE, the apparent molecular mass of PfHMGB factors was approximately 12 kDa, whereas that of HSP70 was approximately 70 kDa. (c) Expression of PfHMGB1 and PfHMGB2 in total lysates prepared from asexual- versus gametocyte-enriched cultures. HSP70 protein expression was used for normalization of sample loading. AS, asexual stages; G, gametocytes.

PfHMGB1 and PfHMGB2, respectively (lanes 3 and 7), slightly smaller than the recombinant His proteins (12.7 kDa and 12.9 kDa, respectively).

Localization of the factors within the nucleus was confirmed by immunofluorescence analysis of unsynchronized parasite cultures containing all asexual stages of *P. falciparum*. Control experiments carried out in the presence of either preimmune sera or the secondary antibody alone showed no signal (data not shown), in contrast to the labeling obtained with both anti-PfHMGB sera coupled with nucleus-specific DAPI staining (Fig. 5). The two proteins appeared to be present mainly in the parasite nuclei, as shown by the merge of DAPI (lanes a and b), and, in contrast to the HSP70 signal, also readily detectable in the cytoplasm, as indicated by the red fluorescence observed in the superposition insert (lane c).

PfHMGB1 and PfHMGB2 factors are expressed differentially in the asexual and gametocyte stages. We compared the expression levels of both factors within total lysates from mixed asexual versus gametocyte cultures by means of Western blotting experiments (Fig. 4c). Each factor was evaluated as detailed in Materials and Methods via HSP70 expression. The level of PfHMGB1 expression was clearly lower in gametocytes than in mixed asexual stages, in contrast to PfHMGB2, whose expression was higher in gametocytes. Indeed, after densitometric quantification of this representative experiment via HSP70 protein normalization (QuantiScan Biosoft 2.1), the normalized values of protein expression were 71%/29% for PfH MGB1 and 27%/73% for PfHMGB2, respectively, when asexual and gametocyte cultures were compared.

We also compared the localizations of both factors in asexual (red immunofluorescence) and gametocyte (green immunofluorescence) stages. As already mentioned, the two PfH MGB factors (Fig. 5, lanes a and b) appeared to be located mainly in the nucleus of the asexual stages (rings, trophozoites, and schizonts), whereas the HSP70 protein (lane c) was also found in the parasite cytoplasm. Surprisingly, in addition to its nuclear localization, PfHMGB2 could also be readily detected within the cytoplasm of different stages (IV and V) of gametocytes (lanes e), as also observed for the HSP70 protein (lane f), whereas PfHMGB1 was associated mainly with the nucleus of gametocytes, as in asexual parasites (Fig. 5, lanes d and a).

DISCUSSION

The two *Plasmodium* HMG proteins belong to the HMGB subfamily and comprise only one HMG box domain, like proteins of plants and several proteins of yeast and *Drosophila*. The parasite HMG box domains are quite similar to those of all eukaryotic organisms and show characteristic residues that are important for DNA binding and bending, as shown for the *Drosophila* HMG-D protein by Murphy and colleagues (47). The presence of Ser-10 and Val-32 (boxed residues in Fig. 1, according to the numbering of HMG-D) allowed us to assign the *Plasmodium* factors to the architectural HMGB family, as also shown by the phylogenetic analysis performed with 159 HMG box domains (see supplemental material S1). Ser-10 forms water-mediated hydrogen bonds with DNA. The hydrophobic residues Val-32 and Met-13 (Fig. 1, residues with asterisks) partially intercalate between two base pairs, introducing two successive kinks into the bound DNA that enhance the more uniform bend associated with the widening of the minor groove. In addition to the HMG box domain, the parasite proteins exhibit a basic extension N terminal to the HMG box domain and apparently no acidic C-terminal tail, as also observed for the yeast NHP6A protein. The HMG box domains bind the DNA minor groove, and in the NHP6A and HMG-D proteins, the basic extension binds in the compressed major groove on the face of the helix opposite the widened minor groove (1, 13), so as to stabilize the HMG box domain-induced bending (39) and consequently facilitate circularization (22, 68).

Since the two *Plasmodium* HMGB factors exhibited only one HMG box domain, we asked whether the *Plasmodium* box was more similar to box A or box B of the metazoan HMGB that encompasses two HMG box domains in tandem. All analyses converge to the same conclusion: the *Plasmodium* HMG box domain more closely resembles box B. When the phylogenetic analysis (see supplemental material S2) was performed with the HMG box domains of PfHMGB1 and PfHMGB2 and of various proteins containing box A and box B, the *Plasmodium* factors clearly clustered with all B boxes. For the human HMG1, it was reported that structure-specific binding to the four-way DNA junction was mediated by the A domain (65) and that box B, flanked by the basic region, displayed a marked DNA recognition activity (69). Hence, the short N-terminal basic domain of the two *Plasmodium* nuclear factors might govern their interaction with distorted DNA and subsequent DNA bending.

In addition, box B of the HMGB of vertebrates was reported to behave as a potent proinflammatory cytokine (64). The tumor necrosis factor (TNF)-stimulating activity was mapped to the KDPNAPKRPPSAFFLFCSEY sequence, corresponding to the first 20 aa of box B in human HMG1 factor, according to the numbering of Li et al. (36). In *Plasmodium*, a domain sharing 75% and 70% identical or strongly similar residues with the TNF-stimulating domain of the human factor was found at the N-terminal position of the HMG box domains of PfHMGB1 and PfHMGB2, respectively. Presently, experiments are under way to analyze whether the two *Plasmodium* factors exhibit TNF-stimulating function.

An automatic three-dimensional structural prediction was performed, and for both *Plasmodium* factors, four α-helices, called 1, 1′, 2, and 3, were predicted (underlined residues of PfHMGB1 and PfHMGB2 of Fig. 1), folding in an L shape in the HMG box domain. HMG box domains have actually exhibited only three α-helices (66), but even if four α-helices were predicted in the two parasite proteins, their positions would be in good agreement with those of *Drosophila melanogaster* HMG-D (PDB file 1qrv [47]), which are indicated (I, II, and III) at the top of Fig. 1. In addition, the sample of the reference template used for nuclear magnetic resonance contained a molecule of β-mercaptoethanol attached to the single cysteine of the protein. The molecule of β-mercaptoethanol, which reduces the affinity of the HMG box domain for the four-way DNA junction, also disrupted the usual first helix (see supplemental material S3). For that reason, helices 1 and 1′ could be regarded as a single, kinked helix.

These computational analyses were concordant in suggesting that the *Plasmodium* proteins were genuine architectural HMGB factors close to box B in being able to bind and bend DNA. In vitro analyses performed with both recombinant pro-

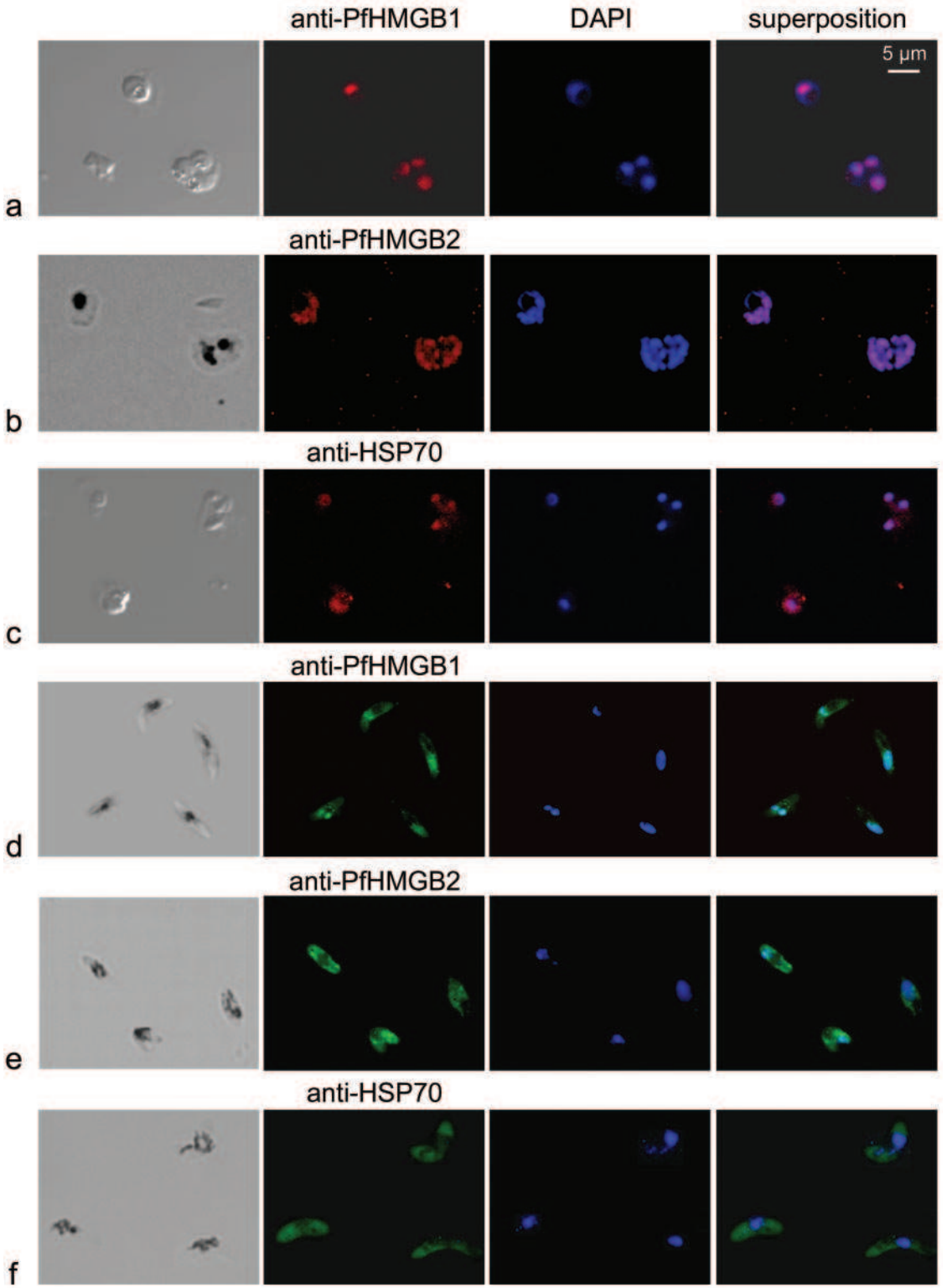


FIG. 5. Immunofluorescence localization of PfHMGB1 and PfHMGB2 in asexual (a, b, and c) and sexual (d, e, and f) stages of *Plasmodium* erythrocytic development. Paraformaldehyde-fixed parasites were labeled with mouse anti-PfHMGB1 and anti-PfHMGB2 antibodies (1:200) and FITC-conjugated anti-mouse IgG (1:100); DNA was stained with DAPI (1:100). Merged fluorescent signals are shown in the "superposition" column. Cells were visualized by phase-contrast (a and c) or transmission (b, d, e, and f) microscopy. Panels: a, trophozoites; b, trophozoite and schizont; c, trophozoites; d to f, gametocytes. Anti-PfHMGB and anti-HSP70 fluorescence is red for panels a to c and green for panels d to f.

teins established that they were indeed able to interact with distorted DNA structures (Fig. 2) and bend linear DNA (Fig. 3), leading to the validation of the computational data (Fig. 1 and supplemental materials S1, S2, and S3). In contrast, EMSA performed with labeled linear DNA binding sites reported to interact specifically with members of the HMGB subfamily comprising the usual TF SOX and SRY (63) gave no detectable retarded complexes (data not shown).

Therefore, it can reasonably be assumed that these architectural HMGB factors might play a role in the remodeling of chromatin. In eukaryotes, one proposed mechanism is that the HMGB nuclear factors might change the nucleosome structure and relax the wrapped DNA so as to enhance the accessibility of the remodeling complexes to chromatin and facilitate interaction of TF with their binding sites (for a review, see reference 62). It has also been observed that the interplay between these factors and the linker histone H1 modulates the balance between alternative conformations of the chromatin, histone H1 enhancing chromatin compaction, in contrast to HMGB. In *Plasmodium*, even though the gene for the H1 linker histone has not yet been annotated, along with 60% of the 5,300 predicted genes, a putative histone H1-like protein might be present and counteract the function of *Plasmodium* HMGB. Indeed, the H1 histones are evolutionarily conserved in metazoans but substantially divergent in protists (51). Some protists appeared to have only a lysine-rich basic protein, whose composition is similar to some of the histone H1-like proteins from eubacteria, animals, and plants (26).

Le Roch et al. reported differential expression of both transcripts. P*fhmgb1* is preferentially expressed during the erythrocytic asexual stages, in contrast to P*fhmgb2*, in which preferential expression occurs in gametocytes (35). Figure 4c shows that expression of the two corresponding proteins is closely related to the level of transcripts and is differentially expressed in mixed asexual and gametocyte stages. The Western blot of cytoplasmic and nuclear extracts prepared from asexual stages (Fig. 4b) and immunofluorescence of asexual and gametocytes stages (Fig. 5a, b, d, and e) revealed that the two factors are localized mainly in the nucleus. Furthermore, PfHMGB2 was clearly detected in the cytoplasm of gametocytes (Fig. 5d and e).

In addition to the differences in the levels of expression and localization within asexual and sexual parasites, these two factors exhibited different affinities when interacting and bending DNA (Fig. 3), PfHMGB2 being less efficient, at least when examined in vitro. All of these results argue in favor of little if any redundancy between the two proteins and in favor of a role in gametocytogenesis.

In summary, a combination of computational and molecular analyses is needed to increase our knowledge of transcriptional regulation of *Plasmodium* genes involved in crucial steps of asexual and sexual erythrocytic development. This report describes the characterization of two *Plasmodium* HMGB factors that appear to exhibit substantial similarity to architectural factors as regards their biological functions, at least when analyzed in vitro for the capacity to interact with distorted DNA and to bend DNA, even though their capacities to do so appeared to be quite different. As in eukaryotes, HMGB factors in *Plasmodium*, since they are highly conserved through evolution, are probably involved in chromatin remodeling. However, even though these proteins were observed in asexual and gametocyte stages, their levels of expression were clearly different, with PfHMGB1 likely implicated in proliferation and PfHMGB2 implicated in differentiation of *Plasmodium*. Much more work will be needed to understand the functions of these two proteins, both as nuclear factors and as cytokines. Invalidation of either the factors or the interaction between the PfHMGB and DNA via gene silencing strategies (20) and antagonists, etc., will increase our knowledge of transcriptional regulation as well as our control of the erythrocytic development of the parasite (proliferation and differentiation). It might provide exciting new therapeutic possibilities. In this regard, it is worthy of note that this type of approach, via disruption of DNA/factor interaction, is currently being evaluated for human cancer therapy with a special focus on HMG proteins (4, 40).

ACKNOWLEDGMENTS

We thank Stephan Soullier for kind help in the phylogenetic studies, Olivier Silvie for the gift of anti-HSP70 serum, and Justine Masson (Inserm U677, Paris) for help with the immunofluorescence facilities. We are grateful to Jennifer Richardson for critical reading of the manuscript.

M.G. and C.B. were financially supported by the Ministère de l'Education Nationale, France. The project was supported by Inserm and UPMC funds to C.V.

REFERENCES

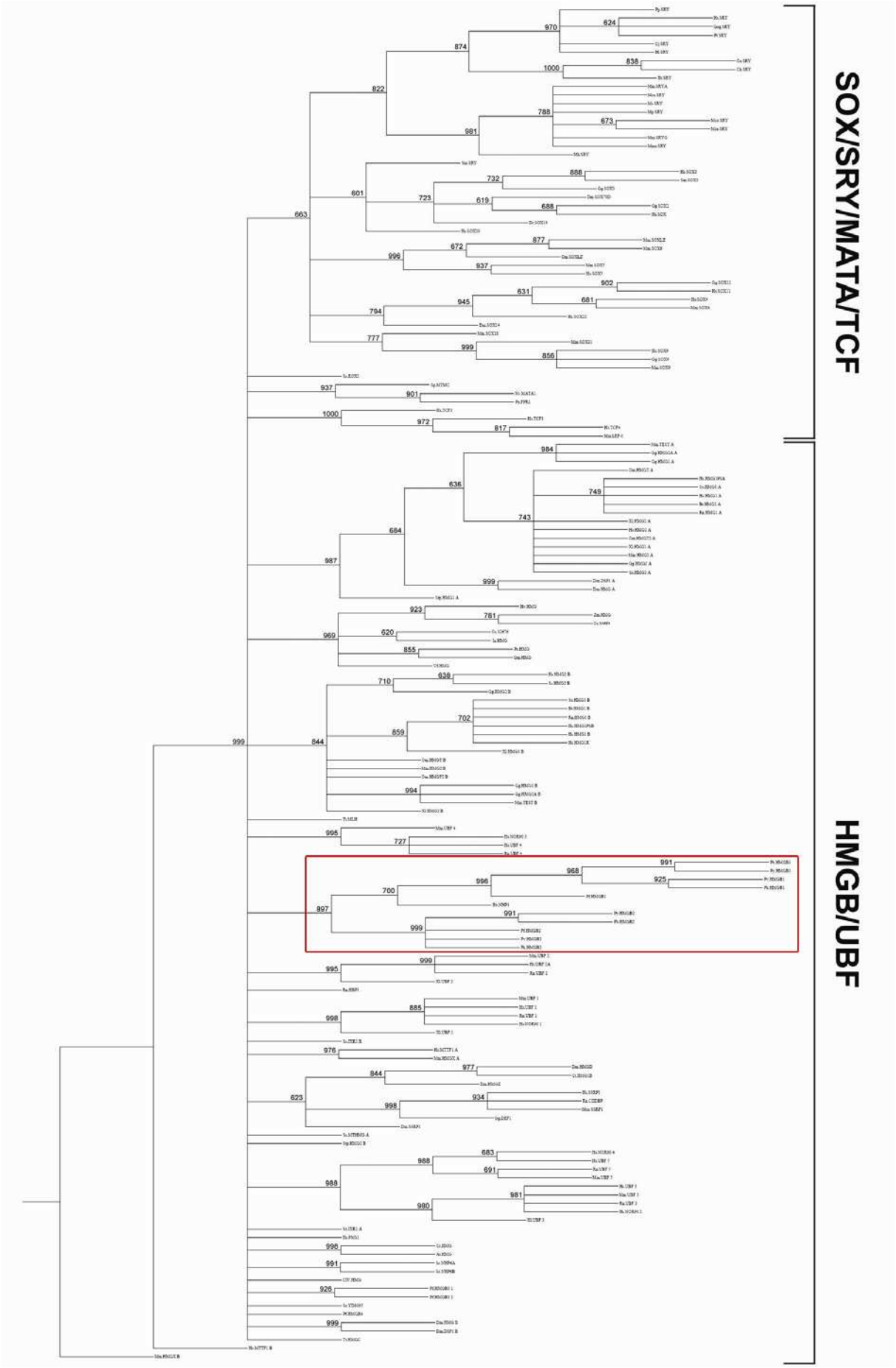
1. **Allain, F. H., Y. M. Yen, J. E. Masse, P. Schultze, T. Dieckmann, R. C. Johnson, and J. Feigon.** 1999. Solution structure of the HMG protein NHP6A and its interaction with DNA reveals the structural determinants for non-sequence-specific binding. EMBO J. **18:**2563–2579.
2. **Aravind, L., L. M. Iyer, T. E. Wellems, and L. H. Miller.** 2003. *Plasmodium* biology: genomic gleanings. Cell **115:**771–785.
3. **Baxevanis, A. D., and D. Landsman.** 1995. The HMG-1 box protein family: classification and functional relationships. Nucleic Acids Res. **23:**1604–1613.
4. **Beckerbauer, L., J. J. Tepe, J. Cullison, R. Reeves, and R. M. Williams.** 2000. FR900482 class of anti-tumor drugs cross-links oncoprotein HMG I/Y to DNA in vivo. Chem. Biol. **7:**805–812.
5. **Bianchi, M. E., M. Beltrame, and G. Paonessa.** 1989. Specific recognition of cruciform DNA by nuclear protein HMG1. Science **243:**1056–1059.
6. **Boschet, C., M. Gissot, S. Briquet, Z. Hamid, C. Claudel-Renard, and C. Vaquero.** 2004. Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 *Plasmodium falciparum* clones. Mol. Biochem. Parasitol. **138:**159–163.
7. **Bustin, M.** 1999. Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. Mol. Cell. Biol. **19:**5237–5246.
8. **Bustin, M.** 2001. Revised nomenclature for high mobility group (HMG) chromosomal proteins. Trends Biochem. Sci. **26:**152–153.
9. **Bustin, M., D. A. Lehn, and D. Landsman.** 1990. Structural features of the HMG chromosomal proteins and their genes. Biochim. Biophys. Acta **1049:** 231–243.
10. **Callebaut, I., K. Prat, E. Meurice, J. P. Mornon, and S. Tomavo.** 2005. Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes. BMC Genomics **6:**100.
11. **Coulson, R. M., N. Hall, and C. A. Ouzounis.** 2004. Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. Genome Res. **14:**1548–1554.
12. **Douguet, D., and G. Labesse.** 2001. Easier threading through web-based comparisons and cross-validations. Bioinformatics **17:**752–753.
13. **Dow, L. K., D. N. Jones, S. A. Wolfe, G. L. Verdine, and M. E. Churchill.** 2000. Structural studies of the high mobility group globular domain and basic tail of HMG-D bound to disulfide cross-linked DNA. Biochemistry **39:**9725–9736.
14. **Dumitriu, I. E., P. Baruah, A. A. Manfredi, M. E. Bianchi, and P. Rovere-Querini.** 2005. HMGB1: guiding immunity from within. Trends Immunol. **26:**381–387.
15. **Erondu, N. E., and J. E. Donelson.** 1992. Differential expression of two mRNAs from a single gene encoding an HMG1-like DNA binding protein of African trypanosomes. Mol. Biochem. Parasitol. **51:**111–118.
16. **Falquet, L., M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann, and A.**

Bairoch. 2002. The PROSITE database, its status in 2002. Nucleic Acids Res. **30:**235–238.
17. **Felsenstein, J.** 1989. PHYLIP—Phylogeny Inference Package (version 3.63). Cladistics **5:**164–166.
18. **Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell.** 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature **419:**498–511.
19. **Gissot, M.** 2005. Etude de la régulation transcriptionnelle des gènes lors du cycle érythrocytaire de *Plasmodium falciparum*. Ph.D. thesis. Université Pierre et Marie Curie—Paris VI, Paris, France.
20. **Gissot, M., S. Briquet, P. Refour, C. Boschet, and C. Vaquero.** 2005. PfMyb1, a *Plasmodium falciparum* transcription factor, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation. J. Mol. Biol. **346:**29–42.
21. **Gnanasekar, M., R. Velusamy, Y. X. He, and K. Ramaswamy.** 2006. Cloning and characterization of a high mobility group box 1 (HMGB1) homologue protein from Schistosoma mansoni. Mol. Biochem. Parasitol. **145:**137–146.
22. **Grasser, K. D., S. H. Teo, K. B. Lee, R. W. Broadhurst, C. Rees, C. H. Hardman, and J. O. Thomas.** 1998. DNA-binding properties of the tandem HMG boxes of high-mobility-group protein 1 (HMG1). Eur. J. Biochem. **253:**787–795.
23. **Hirtzlin, J., P. M. Farber, and R. M. Franklin.** 1994. Isolation of a novel *Plasmodium falciparum* gene encoding a protein homologous to the Tat-binding protein family. Eur. J. Biochem. **226:**673–680.
24. **Horn, P. J., and C. L. Peterson.** 2002. Molecular biology. Chromatin higher order folding—wrapping up transcription. Science **297:**1824–1827.
25. **Horrocks, P., K. Dechering, and M. Lanzer.** 1998. Control of gene expression in *Plasmodium falciparum*. Mol. Biochem. Parasitol. **95:**171–181.
26. **Kasinsky, H. E., J. D. Lewis, J. B. Dacks, and J. Ausio.** 2001. Origin of H1 linker histones. FASEB J. **15:**34–42.
27. **Khochbin, S.** 2001. Histone H1 diversity: bridging regulatory signals to linker histone function. Gene **271:**1–12.
28. **Knapp, S., S. Muller, G. Digilio, T. Bonaldi, M. E. Bianchi, and G. Musco.** 2004. The long acidic tail of high mobility group box 1 (HMGB1) protein forms an extended and flexible structure that interacts with specific residues within and between the HMG boxes. Biochemistry **43:**11992–11997.
29. **Kun, J. F., and R. F. Anders.** 1995. A *Plasmodium falciparum* gene encoding a high mobility group protein box. Mol. Biochem. Parasitol. **71:**249–253.
30. **Labesse, G., and J. Mornon.** 1998. Incremental threading optimization (TITO) to help alignment and modelling of remote homologues. Bioinformatics **14:**206–211.
31. **Langst, G., and P. B. Becker.** 2004. Nucleosome remodeling: one mechanism, many phenomena? Biochim. Biophys. Acta **1677:**58–63.
32. **Lanzer, M., D. de Bruin, and J. V. Ravetch.** 1992. Transcription mapping of a 100 kb locus of *Plasmodium falciparum* identifies an intergenic region in which transcription terminates and reinitiates. EMBO J. **11:**1949–1955.
33. **Lanzer, M., S. P. Wertheimer, D. de Bruin, and J. V. Ravetch.** 1993. *Plasmodium*: control of gene expression in malaria parasites. Exp. Parasitol. **77:**121–128.
34. **Le Roch, K., C. Sestier, D. Dorin, N. Waters, B. Kappes, D. Chakrabarti, L. Meijer, and C. Doerig.** 2000. Activation of a Plasmodium falciparum cdc2-related kinase by heterologous p25 and cyclin H. Functional characterization of a P. falciparum cyclin homologue. J. Biol. Chem. **275:**8952–8958.
35. **Le Roch, K. G., Y. Zhou, P. L. Blair, M. Grainger, J. K. Moch, J. D. Haynes, P. De La Vega, A. A. Holder, S. Batalov, D. J. Carucci, and E. A. Winzeler.** 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science **301:**1503–1508.
36. **Li, J., R. Kokkola, S. Tabibzadeh, R. Yang, M. Ochani, X. Qiang, H. E. Harris, C. J. Czura, H. Wang, L. Ulloa, H. S. Warren, L. L. Moldawer, M. P. Fink, U. Andersson, K. J. Tracey, and H. Yang.** 2003. Structural basis for the proinflammatory cytokine activity of high mobility group box 1. Mol. Med. **9:**37–45.
37. **Li, W. B., D. J. Bzik, H. M. Gu, M. Tanaka, B. A. Fox, and J. Inselburg.** 1989. An enlarged largest subunit of *Plasmodium falciparum* RNA polymerase II defines conserved and variable RNA polymerase domains. Nucleic Acids Res. **17:**9621–9636.
38. **Li, W. B., D. J. Bzik, M. Tanaka, H. M. Gu, B. A. Fox, and J. Inselburg.** 1991. Characterization of the gene encoding the largest subunit of *Plasmodium falciparum* RNA polymerase III. Mol. Biochem. Parasitol. **46:**229–239.
39. **Lnenicek-Allen, M., C. M. Read, and C. Crane-Robinson.** 1996. The DNA bend angle and binding affinity of an HMG box increased by the presence of short terminal arms. Nucleic Acids Res. **24:**1047–1051.
40. **Lotze, M. T., and R. A. DeMarco.** 2003. Dealing with death: HMGB1 as a novel target for cancer therapy. Curr. Opin. Investig. Drugs **4:**1405–1409.
41. **Lotze, M. T., and K. J. Tracey.** 2005. High-mobility group box 1 protein (HMGB1): nuclear weapon in the immune arsenal. Nat. Rev. Immunol. **5:**331–342.
42. **Martinez, E.** 2002. Multi-protein complexes in eukaryotic gene transcription. Plant Mol. Biol. **50:**925–947.
43. **McAndrew, M. B., M. Read, P. F. Sims, and J. E. Hyde.** 1993. Characterisation of the gene encoding an unusually divergent TATA-binding protein (TBP) from the extremely A+T-rich human malaria parasite *Plasmodium falciparum*. Gene **124:**165–171.
44. **Meissner, M., and D. Soldati.** 2005. The transcription machinery and the molecular toolbox to control gene expression in *Toxoplasma gondii* and other protozoan parasites. Microbes Infect. **7:**1376–1384.
45. **Morales, M., E. Onate, M. Imschenetzky, and N. Galanti.**1992 HMG-like chromosomal proteins in *Trypanosoma cruzi*. J. Cell. Biochem. **50:**279–284.
46. **Muller, S., P. Scaffidi, B. Degryse, T. Bonaldi, L. Ronfani, A. Agresti, M. Beltrame, and M. E. Bianchi.** 2001. New EMBO members' review. The double life of HMGB1 chromatin protein: architectural factor and extracellular signal. EMBO J. **20:**4337–4340.
47. **Murphy, F. V. T., R. M. Sweet, and M. E. Churchill.** 1999. The structure of a chromosomal high mobility group protein-DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition. EMBO J. **18:**6610–6618.
48. **Nemeth, A., and G. Langst.** 2004. Chromatin higher order structure: opening up chromatin for transcription. Brief. Funct. Genomics Proteomics **2:**334–343.
49. **Osta, M., L. Gannoun-Zaki, S. Bonnefoy, C. Roy, and H. J. Vial.** 2002. A 24 bp *cis*-acting element essential for the transcriptional activity of *Plasmodium falciparum* CDP-diacylglycerol synthase gene promoter. Mol. Biochem. Parasitol. **121:**87–98.
50. **Paillard, F., G. Sterkers, and C. Vaquero.** 1990. Transcriptional and post-transcriptional regulation of TcR, CD4 and CD8 gene expression during activation of normal human T lymphocytes. EMBO J. **9:**1867–1872.
51. **Parseghian, M. H., and B. A. Hamkalo.** 2001. A compendium of the histone H1 family of somatic subtypes: an elusive cast of characters and their characteristics. Biochem. Cell Biol. **79:**289–304.
52. **Ponnudurai, T., J. H. Meuwissen, A. D. Leeuwenberg, J. P. Verhave, and A. H. Lensen.** 1982. The production of mature gametocytes of *Plasmodium falciparum* in continuous cultures of different isolates infective to mosquitoes. Trans. R. Soc. Trop. Med. Hyg. **76:**242–250.
53. **Ragab, A., and A. Travers.** 2003. HMG-D and histone H1 alter the local accessibility of nucleosomal DNA. Nucleic Acids Res. **31:**7083–7089.
54. **Read, C. M., P. D. Cary, C. Crane-Robinson, P. C. Driscoll, and D. G. Norman.** 1993. Solution structure of a DNA-binding domain from HMG1. Nucleic Acids Res. **21:**3427–3436.
55. **Silvie, O., J. F. Franetich, S. Charrin, M. S. Mueller, A. Siau, M. Bodescot, E. Rubinstein, L. Hannoun, Y. Charoenvit, C. H. Kocken, A. W. Thomas, G. J. Van Gemert, R. W. Sauerwein, M. J. Blackman, R. F. Anders, G. Pluschke, and D. Mazier.** 2004. A role for apical membrane antigen 1 during invasion of hepatocytes by *Plasmodium falciparum* sporozoites. J. Biol. Chem. **279:**9490–9496.
56. **Soullier, S., P. Jay, F. Poulat, J. M. Vanacker, P. Berta, and V. Laudet.** 1999. Diversification pattern of the HMG and SOX family members during evolution. J. Mol. Evol. **48:**517–527.
57. **Su, X. Z., Y. Wu, C. D. Sifri, and T. E. Wellems.** 1996. Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. Nucleic Acids Res. **24:**1574–1575.
58. **Thomas, J. O., and A. A. Travers.** 2001. HMG1 and 2, and related 'architectural' DNA-binding proteins. Trends Biochem. Sci. **26:**167–174.
59. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
60. **Thomsen, M. S., L. Franssen, D. Launholt, P. Fojan, and K. D. Grasser.** 2004. Interactions of the basic N-terminal and the acidic C-terminal domains of the maize chromosomal HMGB1 protein. Biochemistry **43:**8029–8037.
61. **Trager, W., and J. B. Jensen.** 1976. Human malaria parasites in continuous culture. Science **193:**673–675.
62. **Travers, A. A.** 2003. Priming the nucleosome: a role for HMGB proteins? EMBO Rep. **4:**131–136.
63. **van de Wetering, M., M. Oosterwegel, K. van Norren, and H. Clevers.** 1993. Sox-4, an Sry-like HMG box protein, is a transcriptional activator in lymphocytes. EMBO J. **12:**3847–3854.
64. **Wang, H., O. Bloom, M. Zhang, J. M. Vishnubhakat, M. Ombrellino, J. Che, A. Frazier, H. Yang, S. Ivanova, L. Borovikova, K. R. Manogue, E. Faist, E. Abraham, J. Andersson, U. Andersson, P. E. Molina, N. N. Abumrad, A. Sama, and K. J. Tracey.** 1999. HMG-1 as a late mediator of endotoxin lethality in mice. Science **285:**248–251.
65. **Webb, M., and J. O. Thomas.** 1999. Structure-specific binding of the two tandem HMG boxes of HMG1 to four-way junction DNA is mediated by the A domain. J. Mol. Biol. **294:**373–387.
66. **Weir, H. M., P. J. Kraulis, C. S. Hill, A. R. Raine, E. D. Laue, and J. O. Thomas.** 1993. Structure of the HMG box motif in the B-domain of HMG1. EMBO J. **12:**1311–1319.

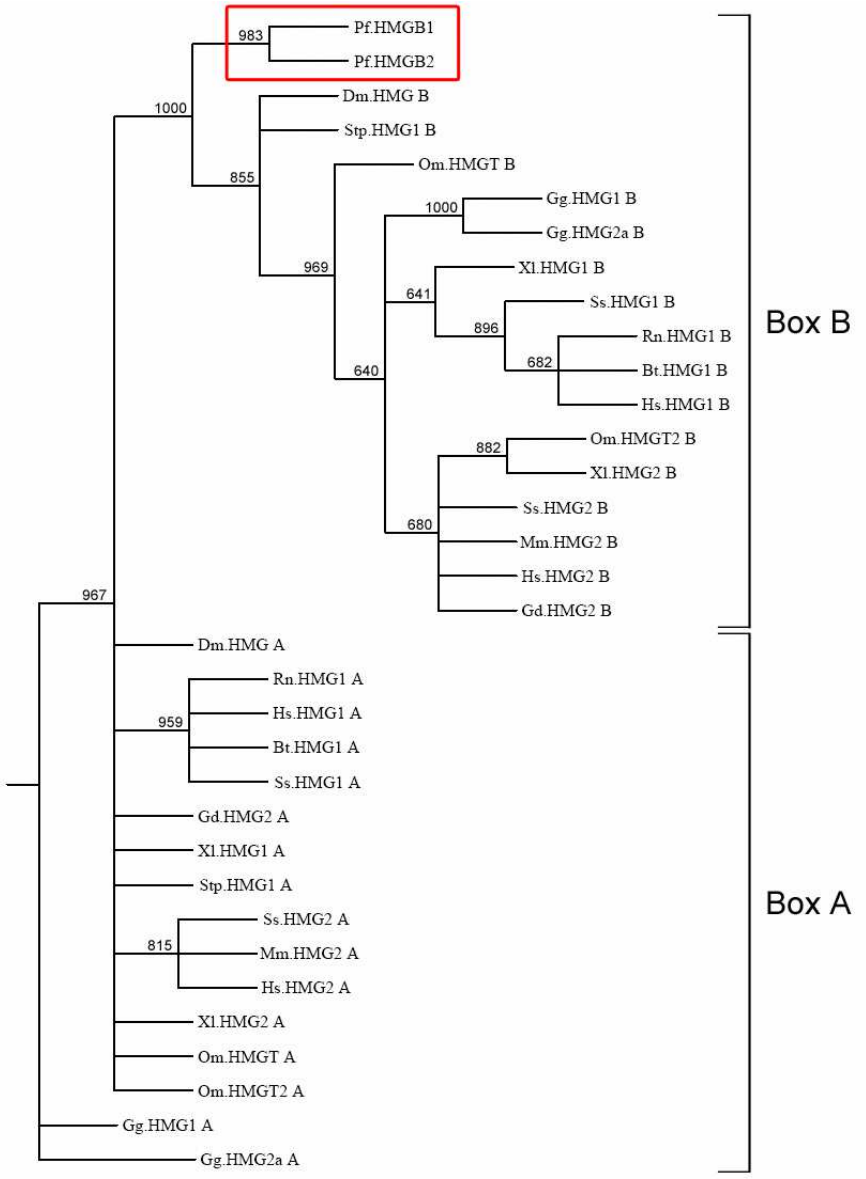
67. **Wu, Q., W. Zhang, K. H. Pwee, and P. P. Kumar.** 2003. Rice HMGB1 protein recognizes DNA structures and bends DNA efficiently. Arch. Biochem. Biophys. **411:**105–111.
68. **Yen, Y. M., B. Wong, and R. C. Johnson.** 1998. Determinants of DNA binding and bending by the *Saccharomyces cerevisiae* high mobility group protein NHP6A that are important for its biological activities. Role of the unique N terminus and putative intercalating methionine. J. Biol. Chem. **273:**4424–4435.
69. **Yoshioka, K., K. Saito, T. Tanabe, A. Yamamoto, Y. Ando, Y. Nakamura, H. Shirakawa, and M. Yoshida.** 1999. Differences in DNA recognition and conformational change activity between boxes A and B in HMG2 protein. Biochemistry **38:**589–595.

Supplementary material 1. Unrooted phylogenetic neighbour-joining tree.

To the hundred HMG sequences extracted from the GenBank data library according to the study of Soullier *et al.* [347] were added the 4 *P. falciparum* HMGB sequences, as well as the orthologs of PfHMGB1 and PfHMGB2 found by sequence homology in the *P. yoelii*, *P. berghei*, *P. vivax* and *P. knowlesi* genomes and the NHP1 sequence of *Babesia bovis*, an A+T-rich apicomplexan parasite like *Plasmodium*. One hundred and fifty-nine HMG-box domains identified with MotifScan were aligned using ClustalW [370]. For proteins containing more than one HMG-box domain, each box was treated separately. The initial full-length alignment of the 159 HMG-box sequences contained 79 sites, all variable. Then regions of the alignment which were equivocally aligned (such as the most N- and C-terminal parts of the alignment), as well as long insertions present in only one or two closely related sequences, were excluded from the analysis and 71 informative sites remained. Programs of the PHYLIP package [113] were then used to perform a reliable unrooted tree. The SEQBOOT program created 1000 replicates of the alignment by boostrapping and the program PROTDIST computed a distance matrix according to the Jones-Taylor-Thornton model for every replicate. The NEIGHBOR program constructed an unrooted tree for every distance matrix with the neighbour-joining method. The CONSENSE program combined all the data in a consensus tree: all the branches supported by bootstrap values under 600 have been collapsed. The consensus tree is therefore visualized by the DRAWGRAM program. Bootstrap values are indicated for every node. The brackets indicate the two subgroups in the HMGB family: the SOX/SRY/MATA/TCF family, whose members are able to bind specific linear DNA sequences and the HMGB/UBF family, whose members interact with high affinity to distorted DNA structures. The subtree clustering the PfHMGB factors and the *B. bovis* NHP1 factor is boxed in red. Abreviations of organism names are as follows: At, *Arabidopsis thaliana*; Bb, *Babesia bovis*; Bt, *Bos taurus*; Ch, *Capra hircus*; CIV, *Chilo iridescent virus*; Cj, *Callithrix jacchus*; Cr, *Catharanthus roseus*; Cs, *Chelydra serpentina*; Ct, *Chironomus tentans*; Dm, *Drosophila melanogaster*; Dr, *Danio rerio*; Gg, *Gallus gallus*; Gm, *Glycine max*; Gog, *Gorilla gorilla*; Hl, *Hylobates lar*; Hs, *Homo sapiens*; Hv, *Hordeum vulgare*; In, *Ipomoea nil*; Mca, *Mus caroli*; Mce, *Mus cervicolor*; Mco, *Mus cookii*; Mh, *Mastomys hildibrantii*; Mma, *Mus macedonicus*; Mm, *Mus musculus*; Mp, *Mus pahari*; Ms, *Mus spretus*; Nc, *Neurospora crassa*; Oa, *Ovis aries*; Om, *Oncorhynchus mykiss*; Os, *Oryza sativa*; Pa, *Podospora anserina*; Pb, *Plasmodium berghei*; Pf, *Plasmodium falciparum*; Pk, *Plasmodium knowlesi*; Pp, *Pongo pygmaeus*; Ps, *Pisum sativum*; Pt, *Pan troglodytes*; Pv, *Plasmodium vivax*; Py, *Plasmodium yoelii*; Rn, *Rattus norvegicus*; Sc, *Saccharomyces cerevisiae*; Sm, *Sminthopsis macroura*; Sp, *Schizosaccharomyces pombe*; Ss, *Sus scrofa*; Stp, *Strongylocentrotus purpuratus*; Tt, *Tetrahymena thermophila*; Vf, *Vicia faba*; Xl, *Xenopus laevis*; Zm, *Zea mays*.

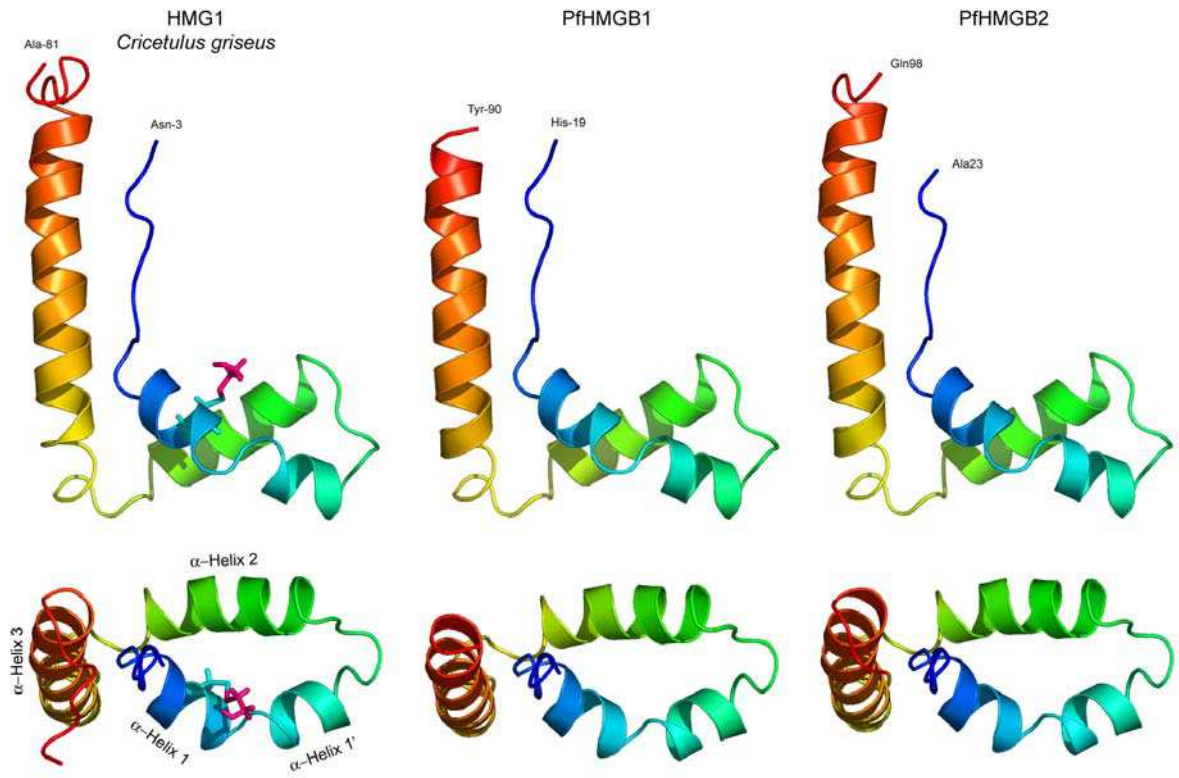

SOX/SRY/MATA/TCF
HMGB/UBF

Supplementary material 2. Unrooted phylogenetic neighbour-joining tree with box A and box B.



The HMG-box domains of PfHMGB1 and PfHMGB2, as well as those of various metazoan proteins containing box A and box B, were aligned with ClustalW. The alignment of the 34 HMG-box domains contained 74 variable sites and was treated as in supplementary data 1. All the branches supported by bootstrap values under 600 have been collapsed. Bootstrap values are indicated for every node. Groups (box A and box B) are indicated by brackets. Abreviations of organism names are as follows: Bt, *Bos taurus*; Dm, *Drosophila melanogaster*; Gd, *Gallus domesticus*; Gg, *Gallus gallus*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Om, *Oncorhynchus mykiss*; Pf, *Plasmodium falciparum*; Rn, *Rattus norvegicus*; Ss, *Sus scrofa*; Stp, *Strongylocentrotus purpuratus*; Xl, *Xenopus laevis*.

Supplementary material 3. Homology modelling of PfHMGB1 and PfHMGB2.



A 3D homology model was built for PfHMGB1 and PfHMGB2 with the @TOME meta-server [@utomatic Threading Optimisation Modelling & Evaluation [94]]. The @TOME meta-server, which submits an amino-acid sequence to six remote servers dedicated to structural predictions and fold recognition, facilitates the recognition of the best 3D template. In the case of PfHMGB1 and PfHMGB2, the best reference template was the box B of the Chinese hamster HMG1 protein [PDB file 1hsn [312]]. Then the TITO program [Tool for Incremental Threading Optimisation [213]] was used to evaluate the compatibility of the amino-acid sequence of the *Plasmodium* factors with this known 3D structure. The sequences from His19 to Tyr90 and from Ala23 to Gln98 were modelled for PfHMGB1 and PfHMGB2, respectively. Whereas HMG-box domains present usually 3 α-helices folded in L-shape, the first α-helix of the reference template (HMG1 of *Cricetulus griseus*) is broken in two sub-helices (α-helices 1 and 1′) because of the presence of a heteroatom of β-mercaptoethanol (in pink) added during the NMR experiments and attached to the single cysteine of the protein (side chain in cyan). The broken first helix is also present in the model of PfHMGB1 and PfHMGB2 (blue helices).

ARTICLE 3

Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 *Plasmodium falciparum* clones.

Charlotte Boschet, Mathieu Gissot, Sylvie Briquet, Zuhal Hamid, Clotilde Claudel-Renard & Catherine Vaquero.

Accepté à la publication dans *Molecular & Biochemical Parasitology*.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular & Biochemical Parasitology 138 (2004) 159–163
MOLECULAR & BIOCHEMICAL PARASITOLOGY

Short communication

Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 *Plasmodium falciparum* clones☆

Charlotte Boschet[a,1], Mathieu Gissot[a,1], Sylvie Briquet[a], Zuhal Hamid[a,2], Clotilde Claudel-Renard[b], Catherine Vaquero[a,*]

[a] *INSERM U511, CHU Pitié-Salpêtrière, 91 boulevard de l'Hôpital, 75013 Paris, France*
[b] *Department of Biochemistry, University of Pretoria, Pretoria, South Africa*

Received 8 May 2004; received in revised form 19 July 2004; accepted 20 July 2004

Keywords: *Plasmodium*; Transcription factor; Expression; Erythrocyte; Development

Two major events of the *Plasmodium falciparum* life cycle occur in the erythrocyte: an asexual multiplication responsible for clinical manifestations and initiation of sexual differentiation responsible for dissemination of the disease via mosquito. These developmental pathways require the coordinated and modulated expression of different sets of genes, involving transcriptional controls, even though post-transcriptional regulation could also be implicated. These gene clusters are therefore likely to share regulatory elements or modules within their promoters [1]. Regulation of transcription is also accomplished by the availability of transcription factors, which is presumably modulated throughout parasite development.

In *Plasmodium* very little is known at the level of *cis*- and *trans*-regulatory elements, and to our knowledge only a few regulatory elements and no clearly defined transcription factors have been reported in the literature. Nevertheless, *Plasmodium* transcriptional machinery appeared to share common features with that of eukaryotes [2]. Moreover, regulatory sequences within the promoters were reported to contain motifs in common with the binding sites of eukaryotic transcription factors [2–5] and our own computer analyses confirmed this statement, including DNA motifs potentially interacting with Myb factors (Boschet and Vaquero, to be published elsewhere).

Herein, we present the first description of a *Plasmodium* transcription factor: a Myb-related protein of the tryptophan cluster family. The Myb proteins are highly conserved in eukaryotes and generally, their characteristic DNA-binding domain (DBD) contains three tandem repeats (R1, R2 and R3) of approximately 50 residues with 3 regularly spaced (18 or 19 amino acids) tryptophan residues [6,7]. They have been shown to bind DNA in a sequence-specific manner and to participate in the regulation of the expression of genes implicated in growth control and differentiation [7]. To identify Myb proteins in the *P. falciparum* genome, more than 200 non-redundant eukaryotic Myb proteins were aligned with MultAlin and ClustalW [8,9] to generate a consensus sequence corresponding to the characteristic DNA-binding domain. This consensus was used as query for the *Plasmodium* database and allowed the annotation of a 414 amino acid-long ORF: PfMyb1. Since PFSCAN [10] identified only one Myb domain (R2 in Fig. 1A), the complete sequence of PfMyb1 was aligned with the DBD of three proteins, DdMybH, DdMyb2 and DdMyb3 [11–13] of the slime mold *Dictyostelium discoideum*, which has an A + T-rich genome (78% overall) like *Plasmodium* (85%). This resulted in the

Abbreviations: aa, amino-acid; DBD, DNA-binding domain; EMSA, electrophoretic mobility shift assay; MRE, Myb regulatory element; NE, nuclear extract; ORF, open reading frame; PCR, polymerase chain reaction; RT, reverse transcription

☆ *Note:* EMBL accession number AJ291747.

* Corresponding author. Tel.: +33 1 40778111; fax: +33 1 45838858.
E-mail address: vaquero@ext.jussieu.fr (C. Vaquero).

[1] The authors wish it to be known that, in their opinion, Charlotte Boschet and Mathieu Gissot should be regarded as joint first authors.

[2] Permanent address: INMO, University of Gezira, Wad Madani, Box 20, Sudan.

0166-6851/$ – see front matter
doi:10.1016/j.molbiopara.2004.07.011

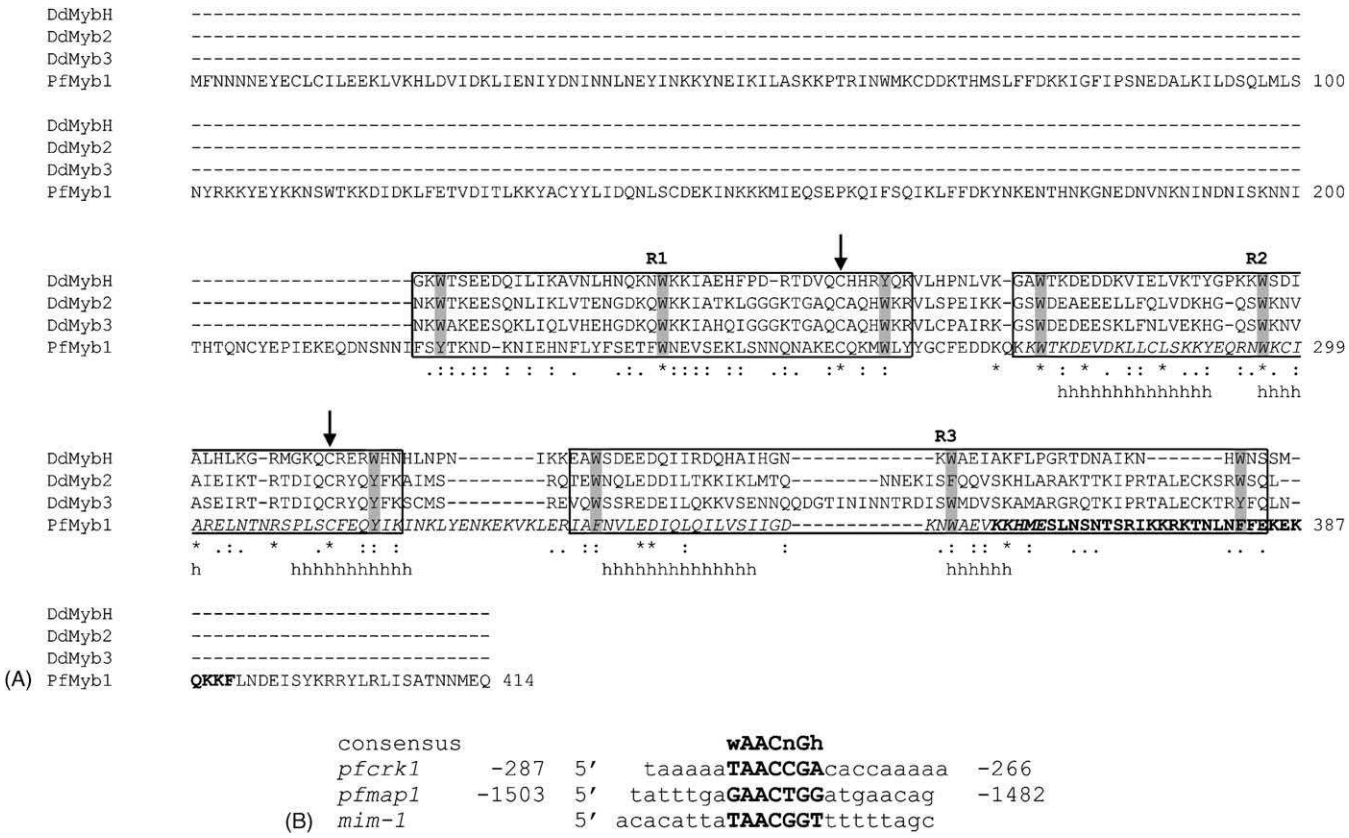


Fig. 1. The DNA-binding domain of the PfMyb1 sequence and three different Myb regulatory elements. (A) The protein consensus of the characteristic DNA-binding domain of the Myb family was obtained with an alignment of eukaryotic Myb selected throughout evolution and allowed the annotation PfMyb1 within the *Plasmodium* chromosome 13. The complete sequence of PfMyb1 is shown as well as an optimal alignment between its putative DNA-binding domain and the three DBD of *Dictyostelium discoideum* (DdMybH, DdMyb2 and DdMyb3; SwissProt accession numbers: P34127, O15816, Q9GUB3, respectively). Identities are indicated by asterisks, strong similarities by two points and weak similarities by one point. The residues W, Y or F that characterize the Myb domains are highlighted in gray and R1, R2 and R3 are boxed. The highly conserved cysteine residues are marked by an arrow. The sequence from Lys 275 to Glu 364 threaded by Meta-Server and TITO [21,35] is indicated in italics and aa predicted to take part in α-helices are marked with an h. A putative nuclear localization sequence, encompassing the unusually long C-terminal part of the third repeat, is shown in bold letters. (B) Putative MRE were localized among others in *Plasmodium pfcrk1* and *pfmap1* gene promoters using MatInspector [36] according to three matrix families (one from the plant and two from the vertebrate subsections). Core and matrix similarity parameters were left by default. These two potential MRE were aligned with a prototype MRE found in the chicken *mim-1* gene promoter reported to be functional [26] and with the consensus sequence of the Myb binding domain. Letters in upper-case represent the MRE and in lower-case the flanking sequences. The positions are numbered from the first ATG of the open reading frames. W: A or T; H: A, C or T; N: A, C, G or T.

identification of three Myb domains (Fig. 1A), located in the C-terminus of the protein as in DdMyb2 and DdMyb3 of *D. discoideum*, and these two factors were reported to be functional [12,13] whereas in most of the Myb proteins reported so far, the DBD is located in the N-terminus. However, PfMyb1 contains imperfect repeats with a tyrosine or a phenylalanine in place of tryptophan and with an aa insertion in R3 as observed in *D. discoideum* (Fig. 1A) as well as in *Saccharomyces cerevisiae* Bas1 [14,15] and *Kluyveromyces lactis* Reb1 [16]. Moreover, a critical cysteine residue, as highly conserved as the tryptophan residues, [17] was also found in R1 and R2 at position 258 and 312, respectively, and was reported to play a role in redox regulation [18–20]. Furthermore, 3D-structure was determined with Meta-Server [21] and it appeared that the sequence from Lys 275 to Glu 364 (Fig. 1A), encompassing the R2 domain and the first part of the R3 domain, could be threaded by homology with the human Myb proto-oncogene protein, whose structure has been determined by X-ray diffraction [22]. Indeed, three α-helices were predicted in R2 and only two in R3 in contrast to the R1 repeat. Nevertheless, in Myb proteins, R2 and R3 constitute a minimal DBD sufficient for sequence-specific DNA binding, while R1 does not appear to engage in specific interactions with DNA [23]. In summary, the diverse computational analyses of PfMyb1 provide evidence for a genuine Myb protein, which is strikingly conserved in all *Plasmodium* species listed in PlasmoDB.

After annotation of the putative transcription factor PfMyb1, the characterization of its cognate RNA transcript was carried out by Northern blotting experiment. A messenger of 2.6 kb was evidenced (data not shown) in *P. falciparum* infected erythrocytes. Our aim was to analyze the differences

occurring at the level of transcription factors that might control expression of transcripts and/or proteins involved in sexual differentiation, a process crucial for the dissemination of the disease via the mosquito. Using a semi-quantitative RT-PCR, the temporal expression of *pfmyb1* transcript (data not shown) was analyzed at different times of erythrocytic development of highly synchronized parasites (rings, early and late trophozoites, as well as early and late schizonts). The profile of expression was determined in two clones of *P. falciparum*, 3D7 and the gametocyte-less F12 derived from 3D7. The major difference in the mRNA profiles resides in a lower abundance of the *pfmyb1* transcript in the ring stage of F12 when compared to 3D7, followed by a sharp increase in early F12 trophozoites. In contrast, the mRNA level was quite similar in both clones in the early and late schizonts after a maximal expression in trophozoites. The 3D7 expression profile is in good agreement with the published transcriptome data of HB3 [24] and 3D7 clones [25].

We then analyzed the presence of the transcription factor in the nuclear extracts of the parasite. The *pfmyb1* ORF has been cloned and the His-tagged recombinant protein was expressed, purified and used as a positive control in Western blots. The PfMyb1 protein was shown to be present in the parasite trophozoite nuclear extracts by a serial immunoprecipitation and Western blotting experiments (Fig. 2A). Briefly, nuclear extracts were immunoprecipitated, using tosylactivated magnetic Dynabeads coated with either anti-PfMyb1 serum (lane 2) or preimmune (lane 3). Immunoprecipitated

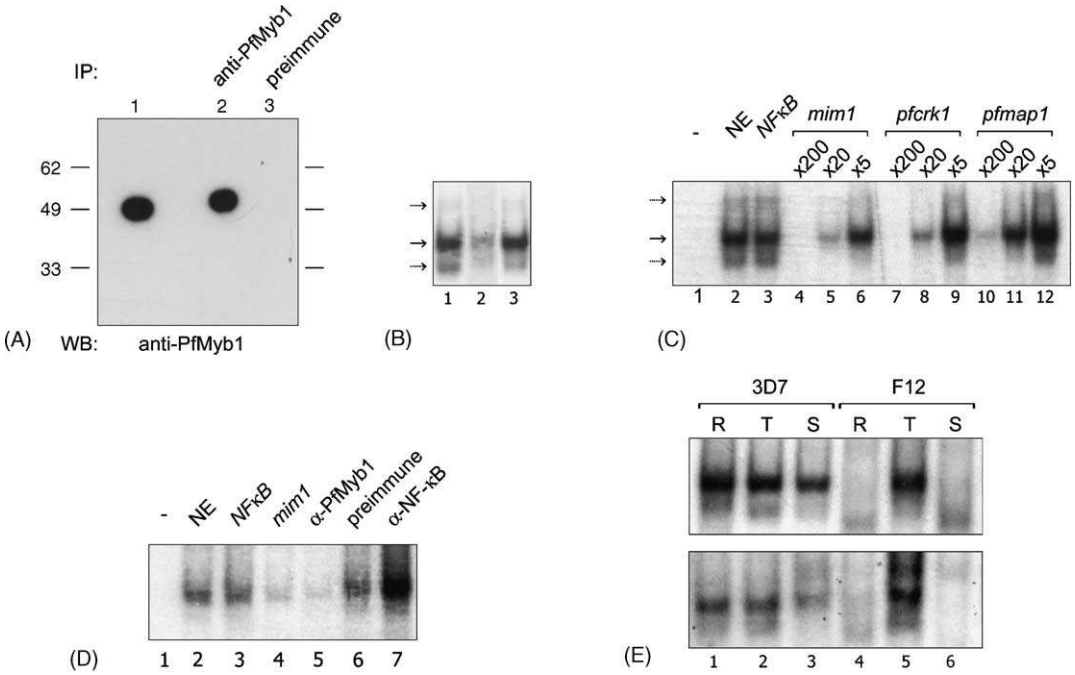


Fig. 2. Functional study of PfMyb1 in *P. falciparum* nuclear extracts. (A) Serial immunoprecipitation and Western blotting experiments. *Plasmodium falciparum* protein extracts were prepared as described in [37]. A polyclonal antipeptide serum was prepared by Eurogentec (Belgium). Two rabbits were immunized with two peptides (RKKYEYKKNSWTKK and KRRYLRLISATNNMEQ), corresponding to amino acids (aa) 103–117 (upstream to the R1 domain) and 400–414 (encompassing the very last C-terminus aa of R3) of the PfMyb1 amino acid sequence, respectively. One hundred micrograms proteins of *P. falciparum* trophozoites were subjected to immunoprecipitation using 2×10^7 magnetic Dynabeads tosylactivated M-280 (Dynal Biotech) coated with 2.5 μg of anti-PfMyb1 serum (lane 2) or preimmune (lane 3) following manufacturer recommendations. Lane 1 stands for PfMyb1 recombinant protein. Immunoprecipitated proteins were run on a 10% SDS-PAGE and subjected, after transfer onto nitrocellulose membrane, to Western blotting using the anti-PfMyb1 serum and revealed by a peroxidase-conjugated anti-rabbit antibody (Sigma). (B–E) EMSA. Nuclear extracts (NE) from 50 mL of red blood cells at 10%–12% parasitaemia infected with rings (1–18 h), trophozoites (20–28 h) and schizonts (32–42 h post-invasion) were prepared as described by Osta et al. [37]. EMSA experiments were conducted with nuclear extracts of 3D7 and F12 clones for 30 min at 4 °C in binding buffer (25 mM HEPES pH 7.5, 50 mM KCl, 1 mM MgCl2, 0.5 mM EDTA, 0.2 mM PMSF, 1 mM DTT, 5% glycerol and 2 μg of poly dI-dC). Addition of 50,000 cpm of labeled probes is followed in a final volume of 10 μl for an additional 30 min. Samples were loaded on pre-run 6% non-denaturing polyacrylamide gel and electrophoresed at 160 V for 2 h at 4 °C in a low-ionic strength TBE buffer (0.25×). (B) Double stranded oligonucleotides (100 ng) of the prototype and putative *Plasmodium* MRE (Fig. 1B) were end-labeled with 40 μCi of [γ-^{32}P] ATP and purified with QIAquick nucleotide removal kit (Qiagen). One micrograms of 3D7 trophozoite nuclear extract was incubated with labeled chicken *mim-1* prototype, *pfmap1* and *pfcrk1* putative *Plasmodium* MRE (lanes 1–3). A bold arrow shows the major complex and dotted arrows indicate two minor complexes. (C) Competitions were carried out as in (B) only with the labeled chicken *mim-1* MRE prototype in presence of a 200-fold molar excess of unlabeled non-specific *NF-κB* (lane 3) and a 200-, 20- and 5-fold excess of specific *mim-1* (lanes 4–6), *pfcrk1* (lanes 7–9) and *pfmap1* (lanes 10–12) oligonucleotides. Controls without and with NE alone are presented in lanes 1 and 2, respectively. (D) Interaction between 3D7 nuclear extract and *mim-1* was assayed in the presence of either 0.5 μl of anti-PfMyb1 polyclonal antibody (0.6 μg/μl, lane 5), or 0.5 μl of the corresponding preimmune serum (lane 6), or 0.5 μl of an irrelevant anti-NF-κB antibody (lane 7). Lanes 1–4 are as in (C). (E) Labeled *mim-1* oligonucleotide was incubated with 1 μg of nuclear extracts of 3D7 and F12 rings (R, lanes 1 and 4), trophozoites (T, lanes 2 and 5) and schizonts (S, lanes 3 and 6). Two independent experiments are presented.

proteins were subjected to SDS-PAGE and Western blotting with the anti-PfMyb1 serum. A single band, whose molecular mass (50 kDa) is slightly higher than the recombinant PfMyb1 (lane 1) was observed (lane 2) in the parasite nuclear extracts. In contrast, when using preimmune coated beads (lane 3) no protein was revealed by the anti-PfMyb1 serum.

In order to activate or repress transcription of particular sets of genes, the transcription factors have to bind *cis*-regulatory DNA elements in a sequence-specific manner. A Myb-related protein was evidenced in the 3D7 trophozoite nuclear extracts (NE) since specific interactions were detected with diverse Myb-regulatory elements (Fig. 1B), such as a prototype MRE (*mim-1* gene promoter) [26] and two putative MRE naturally occurring in the promoters of *Plasmodium pfmap1* and *pfcrk1* genes. These genes were originally reported to be expressed preferentially during erythrocytic asexual and sexual stages, respectively [27,28]. The EMSA profiles (Fig. 2B) appeared quite similar for the three MRE assayed, all giving rise to one major and two minor complexes, even though the magnitude of interaction was different, with its highest level being observed for *mim-1* and *pfcrk1* MRE, in line with the nucleotide sequence of the elements (Fig. 1B). These retarded bands could result from homo-dimerization of the PfMyb1 protein, as well as from interaction with other proteins, which have been shown to interact with the Myb proteins, as the bZip transcription factor C/EBPβ [22] and the poly(ADP-ribose) polymerase (PARP) [29]. The DNA/protein interaction was specific, as shown by competition experiments with unlabeled oligonucleotides corresponding to each MRE (Fig. 2C). The interaction was not competed out by a 200-fold excess of an irrelevant unlabeled *NF-κB* element present in the *IL-2Rα* promoter [30] (lane 3), in contrast to a 200-fold excess of unlabeled relevant *mim-1*, *pfcrk1* and *pfmap1* (lanes 4, 7 and 10, respectively). In Fig. 2D, the specificity of the interaction was verified by using the anti-PfMyb1 antibody (lane 5), which competed out interaction with the labeled probe as efficiently as the *mim-1* competitor (lane 4), in contrast to the pre-immune serum (lane 6) and an irrelevant anti-NF-κB antibody [31] (lane 7). Indeed, anti-PfMyb1 antibody, raised against the C-terminus of the DBD, impaired complex formation in place of generating a super-shifted band as already reported for other transcription factors [30,32]. For all these reasons we think that PfMyb1 is a genuine transcription factor. However, it was not possible, under the experimental conditions assayed so far, to demonstrate any interaction between these MRE and the recombinant protein, whether complete or incomplete, encompassing the three Myb domains. This might be due to inappropriate folding, processing, and/or phosphorylation of the recombinant protein, as has already been reported for other recombinant *P. falciparum* proteins expressed in *E. coli* [33].

Finally, we compared the temporal expression of PfMyb1 protein during 3D7 and F12 erythrocytic development with nuclear extracts prepared from the three main forms: rings, trophozoites and schizonts. EMSA experiments using the three nuclear extracts and labeled *mim-1* oligonucleotide are shown in Fig. 2E and correspond to two representative and independent biological experiments. The major complex was present throughout development of the 3D7 clone, decreasing only slightly up to the schizont stage (lanes 1–3). In contrast, for the F12 clone the major complex, while quite apparent in trophozoites (lane 5), was almost undetectable in rings and schizonts (lanes 4 and 6). A discrepancy between transcript and EMSA profiles was quite apparent, especially in the F12 clone. The messenger was observed in F12 rings, despite absence of Myb/MRE interaction. Moreover, in early and late schizonts of both clones, the *pfmyb1* transcript was present to equal extents, while the DNA/protein interaction was detectable in 3D7 and undetectable in F12 nuclear extracts. The absence of interaction was not due to the degradation or lower concentration of the nuclear extracts used in the EMSA, since the integrity and protein content of all nuclear preparations were verified by SDS-PAGE analysis and Bradford assay (data not shown). Decreased mRNA translation [34] or absence of expression of a co-factor responsible for the formation of EMSA complexes, among other explanations, could account for the absence of retarded band. We are not, however, as yet able to address these questions. Nevertheless, the lack of PfMyb1 or partners in rings and schizonts might participate to the impairment of sexual differentiation towards gametocytes observed in F12 clone.

In summary, by computational annotation we identified, within the genomic DNA of *P. falciparum*, an open reading frame sharing common features with the Myb transcription factors. We assumed that this factor could regulate expression of particular genes since appropriate binding sites were identified within their promoters. In nuclear extracts, a Myb DNA binding activity was observed with a prototype and two putative *Plasmodium* Myb regulatory elements. This interaction was demonstrated to be specific, since it was inhibited by specific competitors and anti-PfMyb1 antibody in bandshift assays. Finally, during erythrocytic development, the bandshift profiles were markedly different in the 3D7 and the gametocyte-less F12 clones, in contrast to the transcript profile. These experiments provide the first evidence of the presence, in *Plasmodium*, of a transcription factor belonging to the highly conserved Myb family that might be implicated in the regulation of gene expression.

Acknowledgements

We thank Pietro Alano for his gift of F12 clone, Leila Gannoun (UMR-CNRS 5539) for the nuclear extract protocol and Alain Israël for the anti-NF-κB antibody. We are grateful to Jennifer Richardson for critical reading of the manuscript, and to Professors M.M. Magzoub (Minister of Higher Education and Scientific Research, Sudan) and Dominique Mazier, as well as Philippe Refour (INSERM U511), for continuous encouragement and support. C.B. and M.G. were financially supported by the Ministère de l'Education Nationale, France,

and Z.H. by the WHO Mediterranean Office and the Ministry of Higher Education and Scientific Research, Sudan. The project was supported by INSERM funds to C.V.

References

[1] Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. Development 1997;124:1851–64.
[2] Horrocks P, Dechering K, Lanzer M. Control of gene expression in *Plasmodium falciparum*. Mol Biochem Parasitol 1998;95:171–81.
[3] Lanzer M, de Bruin D, Ravetch JV. Transcription mapping of a 100 kb locus of *Plasmodium falciparum* identifies an intergenic region in which transcription terminates and reinitiates. EMBO J 1992;11:1949–55.
[4] Lanzer M, Wertheimer SP, de Bruin D, Ravetch JV. *Plasmodium*: control of gene expression in malaria parasites. Exp Parasitol 1993;77:121–8.
[5] Horrocks P, Kilbey BJ. Physical and functional mapping of the transcriptional start sites of *Plasmodium falciparum* proliferating cell nuclear antigen. Mol Biochem Parasitol 1996;82:207–15.
[6] Kanei-Ishii C, Nomura T, Ogata K, et al. Structure and function of the proteins encoded by the *myb* gene family. Curr Top Microbiol Immunol 1996;211:89–98.
[7] Lipsick JS. One billion years of Myb. Oncogene 1996;13:223–35.
[8] Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 1988;16:10881–90.
[9] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.
[10] PFSCAN, http://hits.isb-sib.ch/cgi-bin/PFSCAN.
[11] Stober-Grasser U, Brydolf B, Bin X, Grasser F, Firtel RA, Lipsick JS. The Myb DNA-binding domain is highly conserved in *Dictyostelium discoideum*. Oncogene 1992;7:589–96.
[12] Otsuka H, Van Haastert PJ. A novel Myb homolog initiates *Dictyostelium* development by induction of adenylyl cyclase expression. Genes Dev 1998;12:1738–48.
[13] Guo K, Anjard C, Harwood A, Kim HJ, Newell PC, Gross JD. A myb-related protein required for culmination in *Dictyostelium*. Development 1999;126:2813–22.
[14] Tice-Baldwin K, Fink GR, Arndt KT. BAS1 has a Myb motif and activates HIS4 transcription only in combination with BAS2. Science 1989;246:931–5.
[15] Hovring I, Bostad A, Ording E, Myrset AH, Gabrielsen OS. DNA-binding domain and recognition sequence of the yeast BAS1 protein, a divergent member of the Myb family of transcription factors. J Biol Chem 1994;269:17663–9.
[16] Morrow BE, Ju Q, Warner JR. A bipartite DNA-binding domain in yeast Reb1p. Mol Cell Biol 1993;13:1173–82.
[17] Pinson B, Brendeford EM, Gabrielsen OS, Daignan-Fornier B. Highly conserved features of DNA binding between two divergent members of the myb family of transcription factors. Nucleic Acids Res 2001;29:527–35.
[18] Myrset AH, Bostad A, Jamin N, Lirsac PN, Toma F, Gabrielsen OS. DNA and redox state induced conformational changes in the DNA-binding domain of the Myb oncoprotein. EMBO J 1993;12:4625–33.
[19] Guehmann S, Vorbrueggen G, Kalkbrenner F, Moelling K. Reduction of a conserved Cys is essential for Myb DNA-binding. Nucleic Acids Res 1992;20:2279–86.
[20] Grasser FA, LaMontagne K, Whittaker L, Stohr S, Lipsick JS. A highly conserved cysteine in the v-Myb DNA-binding domain is essential for transformation and transcriptional trans-activation. Oncogene 1992;7:1005–9.
[21] Douguet D, Labesse G. Easier threading through web-based comparisons and cross-validations. Bioinformatics 2001;17:752–3.
[22] Tahirov TH, Sato K, Ichikawa-Iwata E, et al. Mechanism of c-Myb-C/EBP beta cooperation from separated sites on a promoter. Cell 2002;108:57–70.
[23] Ogata K, Morikawa S, Nakamura H, et al. Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. Cell 1994;79:639–48.
[24] Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol 2003;1:5.
[25] Le Roch KG, Zhou Y, Blair PL, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 2003;301:1503–8.
[26] Ness SA, Marknell A, Graf T. The v-myb oncogene product binds to and activates the promyelocyte-specific mim-1 gene. Cell 1989;59:1115–25.
[27] Doerig C, Horrocks P, Coyle J, et al. Pfcrk-1, a developmentally regulated cdc2-related protein kinase of *Plasmodium falciparum*. Mol Biochem Parasitol 1995;70:167–74.
[28] Doerig CM, Parzy D, Langsley G, Horrocks P, Carter R, Doerig CD. A MAP kinase homologue from the human malaria parasite, *Plasmodium falciparum*. Gene 1996;177:1–6.
[29] Cervellera MN, Sala A. Poly(ADP-ribose) polymerase is a B-MYB coactivator. J Biol Chem 2000;275:10692–6.
[30] Galio L, Briquet S, Cot S, Guillet JG, Vaquero C. Analysis of interactions between huGATA-3 transcription factor and three GATA regulatory elements of HIV-1 long terminal repeat, by surface plasmon resonance. Anal Biochem 1997;253:70–7.
[31] Blank V, Kourilsky P, Israel A. Cytoplasmic retention, DNA binding and processing of the NF-kappa B p50 precursor are controlled by a small region in its C-terminus. EMBO J 1991;10:4159–67.
[32] Markle D, Das S, Ward SV, Samuel CE. Functional analysis of the KCS-like element of the interferon-inducible RNA-specific adenosine deaminase ADAR1 promoter. Gene 2003;304:143–9.
[33] Pandey KC, Singh S, Pattnaik P, et al. Bacterially expressed and refolded receptor binding domain of *Plasmodium falciparum* EBA-175 elicits invasion inhibitory antibodies. Mol Biochem Parasitol 2002;123:23–33.
[34] Dechering KJ, Thompson J, Dodemont HJ, Eling W, Konings RN. Developmentally regulated expression of pfs16, a marker for sexual differentiation of the human malaria parasite *Plasmodium falciparum*. Mol Biochem Parasitol 1997;89:235–44.
[35] Labesse G, Mornon J. Incremental threading optimization (TITO) to help alignment and modelling of remote homologues. Bioinformatics 1998;14:206–11.
[36] Quandt K, Frech K, Karas H, Wingender E, Werner T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 1995;23:4878–84.
[37] Osta M, Gannoun-Zaki L, Bonnefoy S, Roy C, Vial HJ. A 24 bp *cis*-acting element essential for the transcriptional activity of *Plasmodium falciparum* CDP-diacylglycerol synthase gene promoter. Mol Biochem Parasitol 2002;121:87–98.

ARTICLE 4

PfMyb1, a *Plasmodium falciparum* transcription factor, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation.

Mathieu Gissot, Philippe Refour, Sylvie Briquet, Charlotte Boschet, Stéphane Coupé, Dominique Mazier & Catherine Vaquero.

Accepté à la publication dans *Journal of Molecular Biology*.

doi:10.1016/j.jmb.2004.11.045 J. Mol. Biol. (2005) 346, 29–42

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

PfMyb1, a *Plasmodium falciparum* Transcription Factor, is Required for Intra-erythrocytic Growth and Controls Key Genes for Cell Cycle Regulation

Mathieu Gissot, Sylvie Briquet, Philippe Refour, Charlotte Boschet and Catherine Vaquero*

INSERM U511, CHU Pitié-Salpêtrière, 91 boulevard de l'Hôpital, 75013 Paris France

During the complex life cycle of *Plasmodium falciparum*, divided between mosquito and human hosts, the regulation of morphologic changes implies a fine control of transcriptional regulation. Transcriptional control, however, and in particular its molecular actors, transcription factors and regulatory motifs, are as yet poorly described in *Plasmodium*. In order to decipher the molecular mechanisms implicated in transcriptional regulation, a transcription factor belonging to the tryptophan cluster family was studied. In a previous work, the PfMyb1 protein, contained in nuclear extracts, was shown to have DNA binding activity and to interact specifically with *myb* regulatory elements. We used long *pfmyb1* double-stranded RNA (dsRNA) to interfere with the cognate messenger expression. Parasite cultures treated with *pfmyb1* dsRNA exhibited a 40% growth inhibition when compared with either untreated cultures or cultures treated with unrelated dsRNA, and parasite mortality occurred during trophozoite to schizont transition. In addition, the *pfmyb1* transcript and protein decreased by as much as 80% in treated trophozoite cultures at the time of their maximum expression. The global effect of this partial loss of transcript and protein was investigated using a thematic DNA microarray encompassing genes involved in signal transduction, cell cycle and transcriptional regulation. SAM software enabled us to identify several genes that were differentially expressed and probably directly or indirectly under the control of PfMyb1. Using chromatin immunoprecipitation, we demonstrated that PfMyb1 binds, within the parasite nuclei, to several promoters and therefore participates directly in the transcriptional regulation of the corresponding genes. This study provides the first evidence of a regulation network involving a *Plasmodium* transcription factor.

Keywords: Plasmodium; gene regulation; transcription; transcription factor; PfMyb1

*Corresponding author

Introduction

Plasmodium is responsible for 1.5–2.7 million deaths annually,[1] mostly among children and pregnant women. Of the four species of *Plasmodium* infecting humans, *Plasmodium falciparum* causes the highest morbidity and mortality. As global efforts to eradicate malaria have failed, there is an urgent need to decipher the biology of *Plasmodium* and in particular the mechanisms of gene regulation that govern its developmental cycle, so as to propose novel strategies to fight malaria.[2]

The cell cycle of *Plasmodium* between vertebrate

Present address: P. Refour, Department of Immunology and Infectious Diseases, Harvard School of Public Health, 665 Huntington Ave., Boston, MA 02115, USA.

Abbreviations used: asRNA, anti-sense RNA; cDNA, DNA complementary to RNA; ChIP, chromatin immunoprecipitation; dsRNA, double-stranded RNA; DNase, deoxyribonuclease; dNTP, deoxyribonucleoside triphosphate; MRE, Myb regulatory element; PCR, polymerase chain reaction; qPCR, real-time quantitative PCR; RNase, ribonuclease; RT, reverse transcription; sRNA, sense RNA.

E-mail address of the corresponding author: vaquero@ext.jussieu.fr

0022-2836/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.

and mosquito hosts implies a high degree of adaptation and strict control of the cellular machinery to correspond precisely to the state of differentiation in the host. These different developmental stages require coordinated modulation of expression of distinct sets of genes, which could be achieved by transcriptional and/or post-transcriptional controls. In *Plasmodium*, as in all eukaryotes, gene expression is governed at the level of transcription by the interaction of elements within promoters acting *in cis* (DNA regulatory elements) and elements acting *in trans* (transcription factors) whose availability is modulated during cellular development. The fine-tuning of transcriptional regulation resides in the interplay between the *cis* and *trans* elements throughout *P. falciparum* development. Little, however, is currently known about these two actors in *Plasmodium*.

The study of *Plasmodium* chromosomes 2 and 3,[3,4] and recently of the complete 3D7 clone sequence,[5] has revealed a genomic organization composed of rather short intergenic sequences, encompassing the basal and distal promoter sequences,[6] between successive open reading frames. The messengers thus far investigated, while few in number, appear to resemble eukaryotic transcripts. They are capped and bear canonic poly-adenylation signals.[7] Moreover, the nucleosome organization,[8] the structure of the promoters and the general transcription factors[9–13] are quite similar to those of other eukaryotes. The regulatory sequences within the promoters appear to contain motifs similar to several binding sites encountered in eukaryotes.[6,14,15] Nevertheless, the majority of the regulatory elements reported so far have not been linked to any known transcription factor[16–18] or found to bind to unknown nuclear factors.[19]

In order to decipher the molecular mechanisms implicated in transcriptional regulation during the erythrocytic development, we have identified, among diverse families of transcription factors, a Myb-related protein belonging to the tryptophan cluster family. In eukaryotes, the Myb proteins are highly conserved throughout evolution and generally contain three repeats of approximately 50 residues with three regularly spaced tryptophan residues.[20,21] Indeed, a large number of Myb-related proteins have been shown to bind DNA in a sequence-specific manner and to participate in the regulation of the expression of genes implicated in growth control and differentiation.[21] PfMyb1 was the first transcription factor not belonging to general transcription factors to be cloned and analyzed in *P. falciparum*.[22] This study showed that *pfmyb1* mRNA reached its highest level at the trophozoite stage, at least as regards the 3D7 clone. In *in vitro* experiments, the PfMyb1 protein present in nuclear extracts was able to bind specifically to a prototype *myb*-regulatory element (MRE) from the chicken *mim-1* gene promoter[23] and two putative MRE annotated within the *Plasmodium* promoters,[22] suggesting that PfMyb1 can be considered to be a Myb transcription factor.

We took advantage of long *pfmyb1* double-stranded RNA (dsRNA) to interfere with expression of the cognate messenger and to address the role of this transcription factor during the erythrocytic cycle. In particular, we investigated the consequences of decrease in *pfmyb1* transcript and protein on parasite growth and transcriptional regulation of *P. falciparum* genes.

Results

pfmyb1 dsRNA alters parasite growth and trophozoite to schizont transition

We investigated the global effect of double-stranded RNA corresponding to part of the *pfmyb1* sequence on *P. falciparum* growth when added to highly synchronized parasite cultures during a 48 hour assay. Parasites were considered to be synchronized after three successive sorbitol treatments leading to an enrichment in rings of at least 95%. In young ring cultures, *pfmyb1* dsRNA or dsRNA of an unrelated gene (HIV-1 *gp120* AN: **K02013**) was added (25 μg per ml of culture) and parasitaemia was evaluated using three different methods (Figure 1). As shown in Figure 1(a), addition of unrelated dsRNA to culture media (line 2) did not affect asexual growth when compared with control cultures (line 1), i.e. parasites grown in media supplemented with identical volumes of annealing buffer. In contrast, a statistically significant decrease in parasitaemia (approximately 40%) was observed in cultures treated with *pfmyb1* dsRNA using manual counting (Figure 1(a), line 3) or a [^{3}H]hypoxanthine assay (Figure 1(a), line 4). The result obtained by manual counting (line 3) corresponds to the average of nine experiments performed in triplicate, while the result obtained by [^{3}H]hypoxanthine assay (line 4) corresponds to two experiments carried out in triplicate. Finally, using a flow-cytometry assay, a similar degree of growth inhibition (40%) was observed after treatment with *pfmyb1* dsRNA as compared with controls treated with single-stranded sense or anti-sense *pfmyb1* RNA (data not shown). Thus three different experimental approaches used for evaluating the inhibition of growth upon addition of dsRNA specific for *pfmyb1* gave essentially the same results (Figure 1).

We investigated the effect of *pfmyb1* dsRNA on trophozoite to schizont transition (Figure 1(b)). During the erythrocytic cycle, smears were performed every four hours and the number of parasites in trophozoite (Figure 1(b), lines 1 and 2) and in late schizont stage (Figure 1(b), lines 3 and 4) was evaluated. The parasitaemia observed in the trophozoite cultures was quite similar in the control (Figure 1(b), line 1) and *pfmyb1* dsRNA-treated cultures (Figure 1(b), line 2), in contrast to that of late schizonts showing a significant decrease in *pfmyb1* dsRNA-treated parasites (Figure 1(b), line 4 *versus* line 3). Moreover, the growth inhibition

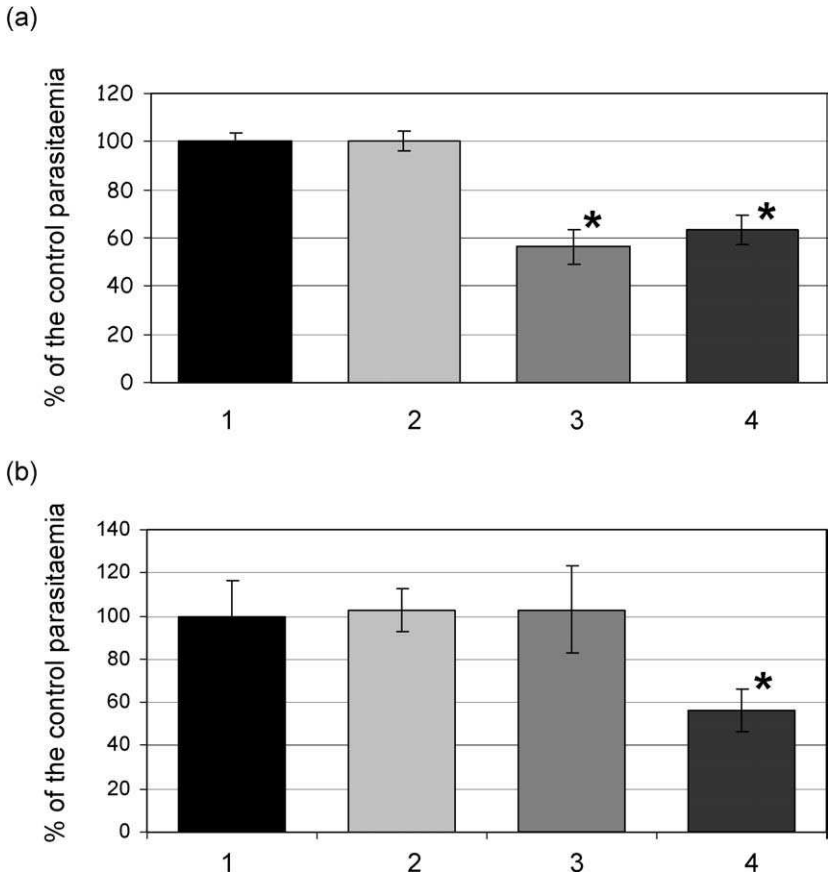


Figure 1. Parasite growth and death evaluation during trophozoite to schizont transition. (a) Evaluation of parasite growth 48 hours after addition of dsRNA to the medium. Parasitaemia was either counted on GIEMSA-stained blood smears (1–3) or monitored by ^{3}H-Hypoxanthine uptake (1 and 4). Parasitaemia was evaluated in untreated culture (1), unrelated dsRNA (2) and *pfmyb1* dsRNA treated culture (3 and 4). Asterisks indicate statistically significant values (p-value <0.0001). Parasitaemia was evaluated in nine independent experiments on Giemsa-stained blood smears and two independent experiments as regards [^{3}H]hypoxanthine uptake assay. (b) Parasitaemia of highly synchronized cultures was determined by counting at late trophozoite (1 and 2) and at late schizont stage (3 and 4) in untreated culture (1 and 3) and *pfmyb1* dsRNA-treated culture (2 and 4). Asterisks are as in (a). The experiment was repeated four times in triplicate. The results are expressed as a percentage of the control cultures.

(Figure 1(a)) and the decrease in schizont number (Figure 1(b)) were of an essentially similar magnitude.

Decreased expression of *pfmyb1* messenger is observed in *pfmyb1* dsRNA-treated cultures

The expression of the *pfmyb1* transcript was analysed at the trophozoite stage, at the peak of expression in the erythrocytic cycle, in cultures treated with *pfmyb1* or *gp120* dsRNA, or left untreated, using a semi-quantitative (Figure 2(a) and (b)) and real-time quantitative reverse transcription (RT)-PCR (Figure 2(b)). For each set of experiments, PCR control reactions were performed in the absence of RNA or RT (data not presented), as well as on the MAL13P1.76 gene with two primers encompassing an intron (Figure 2(a) left panel). These data demonstrated the absence of DNA contamination in the RNA preparations. First, by semi-quantitative RT-PCR (Figure 2(a), right and left panel), we showed that the expression of the *pfmyb1* transcript was decreased in *pfmyb1* dsRNA-treated parasites when compared with control parasites, while expression of the MAL13P1.76 transcript was unaltered. After normalization against ribosomal 18 S RNA, a 75% decrease in the *pfmyb1* transcript was observed in *pfmyb1* dsRNA-treated cultures (Figure 2(b), lines 3 and 6) when compared with control cultures (lines 1 and 4) or cultures treated with unrelated dsRNA (lines 2 and 5), using semi-quantitative (lines 1–3) or real-time quantitative RT-PCR (RT-qPCR) (lines 4–6). RT-qPCR was performed using a multiplex assay amplifying *pfmyb1* and the 18 S reference in the same experiment. For the experiments presented herein, different couples of primers (see Materials and Methods) were used, all yielding similar results.

Decreased expression of PfMyb1 protein is observed in *pfmyb1* dsRNA-treated culture

In order to follow the effect of *pfmyb1* dsRNA treatment on PfMyb1 protein expression, serial immuno-precipitation and Western blotting experiment were performed. Protein extracts from *P. falciparum* trophozoites were immuno-precipitated using magnetic beads coated with anti-PfMyb1 serum. Immuno-precipitated proteins were subjected to SDS-PAGE followed by Western blotting with the anti-PfMyb1 serum. No bands were detected after immuno-precipitation of lysates with either preimmune serum or uncoated beads and Western blotting with the anti-PfMyb1 serum (data not shown). The abundance of the PfMyb1 protein was markedly decreased in *pfmyb1* dsRNA-treated parasites (Figure 3(a), lower panel) as confirmed by normalization with PfHsp70 (Figure 3(a), upper panel). Indeed, a reduction of approximately 80% in the PfMyb1 protein was observed in protein extracts from trophozoite cultures treated with *pfmyb1* dsRNA when

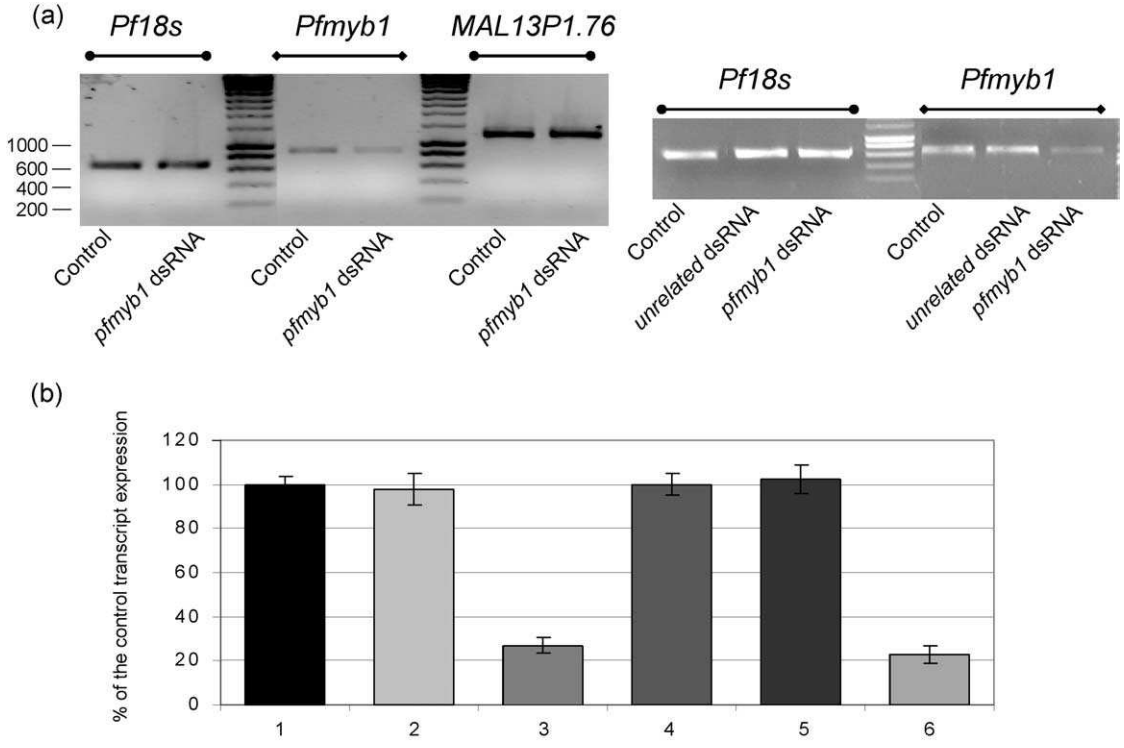


Figure 2. Analysis and quantification of *pfmyb1* mRNA expression. (a) Amplicons of *pfmyb1*, 18 S ribosomal RNA and MAL13P1.76 (indicated on the top) were obtained after semi-quantitative RT-PCR on DNase-treated total RNA prepared from either untreated culture (control), irrelevant *gp120* dsRNA (unrelated dsRNA) or *pfmyb1* dsRNA (*pfmyb1* dsRNA) treated culture. Two independent experiments are presented. Markers are indicated on the left side of the Figure. (b) Graphical representation of the *pfmyb1* amplified fragment intensities after normalization with 18 S ribosomal RNA. Semi-quantitative (1–3) and real time quantitative RT-PCR (4–6) were performed on samples from either untreated culture (1 and 4), or cultures treated with irrelevant (2 and 5) or *pfmyb1* dsRNA (3 and 6). The results are expressed as a percentage of the control transcript expression in three independent experiments.

compared to untreated cultures (Figure 3(b), lane 2 *versus* lane 1), while no such reduction was observed in controls treated with unrelated dsRNA (lane 3).

Genes are differentially expressed in *pfmyb1* dsRNA-treated culture

We then decided to investigate the effect of the partial loss of the PfMyb1 transcript and protein on the global expression profile, by comparing untreated culture with cultures treated with either *pfmyb1* dsRNA or irrelevant dsRNA at a time when the expression of the *pfmyb1* transcript was highest, i.e. the trophozoite stage.

With the use of a thematic microarray fully described by Gissot *et al.*[24] and comprising 180 PCR products printed in duplicate, among which 153 amplified from the *P. falciparum* genomic sequences, we monitored the steady state of the corresponding transcripts. The genes were selected for their homology to known proteins previously described in eukaryotes to be involved in genetic regulation from DNA to proteins. The microarray quality controls were performed as detailed by Gissot *et al.*[24] Briefly, for each growth condition, at least two different total RNA preparations from independent cultures were used for synthesis of cDNA radiolabelled with ^{3}H or ^{35}S radioisotope, and for subsequent hybridization of two sets of arrays. Each raw signal was normalized twice, first with the Alien cDNAs in order to normalize for the efficiency of each reverse transcription and second with 3D7 genomic DNA to normalize for the amount of each cDNA used for the hybridization. Statistically significant differences in gene expression were monitored using the Statistical Analysis for Microarrays (SAM) program.[25] Among the 153 *P. falciparum* genes printed on the microarray, 139 genes gave signals at least twice as high as background (see Supplementary Data 1 and 2).

When untreated controls were compared with HIV-1 *gp120* dsRNA-treated culture, no differences in the relative expression level of any gene were detected using the SAM program (see Supplementary Data 2). In contrast, upon comparison of *pfmyb1* dsRNA treated and untreated parasites, 11 genes were identified as statistically different and the 128 remaining genes as similar. Among the 11 genes, one was up-regulated and ten down-regulated, as shown in the SAM output plot presented in Figure 4.

The differential expression of these 11 genes was

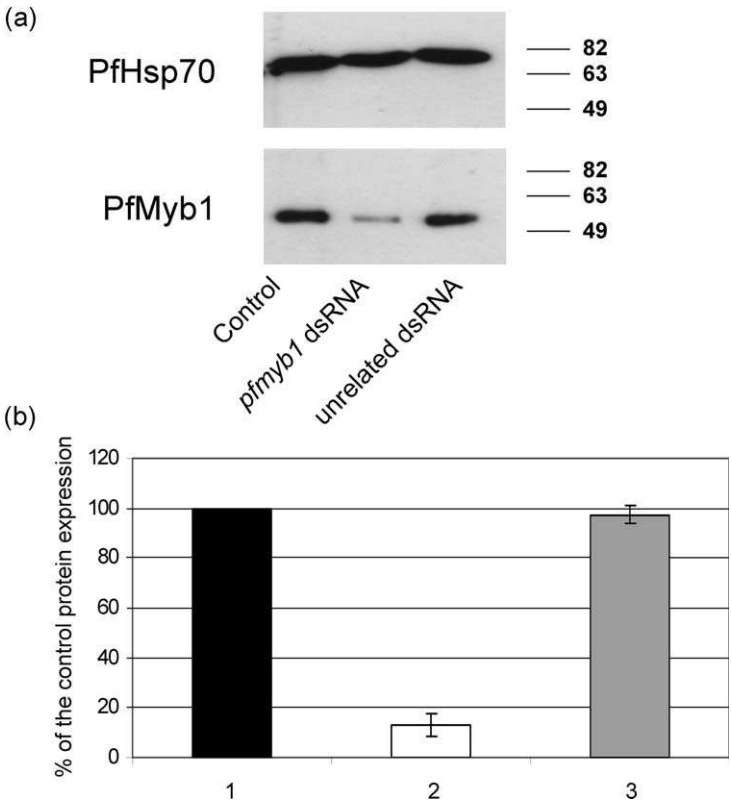


Figure 3. Analysis and quantification of PfMyb1 protein expression. (a) Serial immunoprecipitation and Western blotting of *P. falciparum* PfMyb1. The protein extracts of trophozoite parasites were first subjected to immuno-precipitation using beads coated with an anti-PfMyb1 serum followed by Western blot analysis with the same anti-PfMyb1 serum (lower panel). Concomitantly a Western blot analysis with an anti-PfHSP70 serum was performed (upper panel). The protein markers are indicated on the right of the Figure. Samples from either untreated culture (control), or cultures treated with irrelevant (unrelated dsRNA) or *pfmyb1* dsRNA (*pfmyb1* dsRNA) were studied. (b) Graphical representation of the PfMyb1 protein intensities after normalization with the intensities of the PfHSP70 protein: Lane 1, untreated culture; lane 3, irrelevant dsRNA-treated culture; lane 2, *pfmyb1* dsRNA (*pfmyb1* dsRNA) treated culture. One representative experiment among three is presented in (a). The results are expressed as for Figure 2.

verified by real-time quantitative RT-PCR using the 18 S gene as reference. Using two independent biological samples, three (numbered 2, 3 and 4 in Figure 4) of the ten down-regulated genes did not give sufficiently consistent results to be taken into account. The results of the RT-qPCR (filled bars) and microarray (open bars) analyses, for the genes whose expression was modulated by *pfmyb1* dsRNA (seven down and one up-regulated gene, also listed in Table 1) and diverse additional genes as controls of the experiments, are compared in Figure 5.

Indeed, several genes whose expression profile was similar in control and *pfmyb1* dsRNA-treated parasites were selected and their expression compared by DNA microarray or RT-qPCR experimental approaches. PfHSP70 (PF08_0054) and MAL13P1.76 were shown to be equally expressed in treated and untreated cultures when analysed by either experimental approach. This was also the case for PF10_0233, PFD0465w and MAL8P1.154 genes, which showed slight but detectable homology with the *pfmyb1* sequence used to generate *pfmyb1* dsRNA by using the BLASTN tool on the PlasmoDB web site, as well as for *pfmyb2* (PF10_0327) and *pfmyb3* (PF10_0143), belonging to the tryptophan cluster family of transcription factors. Moreover, the RT-qPCR also confirmed the results obtained for the seven down and one up-regulated gene, attesting to the reliability of the microarray data (Figure 5 and Table 1).

Finally, the over-expression of the *pfpk5* transcript in *pfmyb1* dsRNA-treated samples was also established at the level of protein expression (Figure 6). Western blotting experiments on whole trophozoite protein extracts showed an increased representation of PfPk5 protein in *pfmyb1* dsRNA-treated samples as compared with untreated samples, in good agreement with the transcript data.

PfMyb1 is associated with the promoter of genes differentially expressed using ChIP experiments

In order to determine whether, within the parasite nuclei, the PfMyb1 protein could interact with promoters of genes listed in Table 1, and therefore could be involved in the regulation of the targeted genes, chromatin immuno-precipitation (ChIP) experiments were carried out. PfMyb1 protein present in nuclear extracts has been shown to interact with Myb-related elements (MRE).[22] A computational analysis was performed on the promoters of the genes listed in Table 1 to identify these MRE, and revealed that among the eight promoters only that of PGK (PFI1105w) did not contain any putative MRE. The PfMyb1–DNA complexes were cross-linked within the parasite nuclei and subsequently immuno-precipitated by using the anti-PfMyb1-coated beads. No cross-linked PfMyb1–DNA complexes were immuno-precipitated using either beads alone or beads coated with preimmune serum (data not shown).

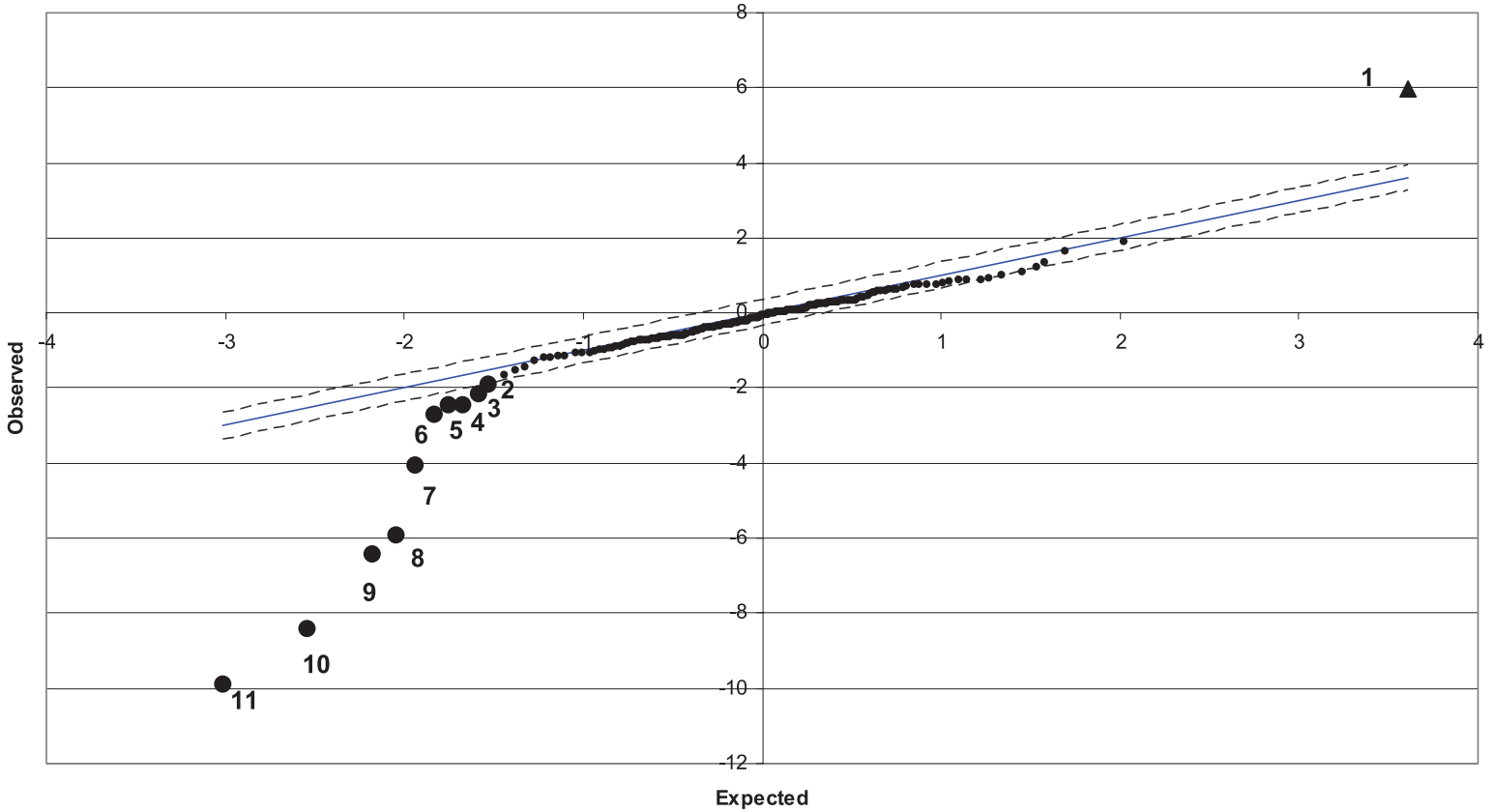


Figure 4. Graphical output of SAM software. Genes are plotted where the *x*-axis represents the expected distribution of each PCR product intensity and the *y*-axis observed distribution of each PCR product intensity. Genes identified by SAM as up-regulated in *pfmyb1* dsRNA-treated samples or as down-regulated in *pfmyb1* dsRNA-treated samples are represented by triangles or circles, respectively, and are as follows: (1) PfPk5 (MAL13P1.279); (2) PF14_0366; (3) PFC0385c; (4) PFE0420c; (5) PfPGK (PFI1105w); (6) PfPk2 (PFL1885c); (7) PfTBP (PFL2345c); (8) PfPCNA1 (PFL1285c); (9) PP1-like (PF14_0224); (10) PfHistone H3 (MAL6P1.248); and (11) PfHistone H2A (MAL6P1.249).

Table 1. List of SAM-identified genes verified by qPCR

Gene name	AN[a]	Putative function	Number[b]	Microarray results[c]
PGK	PFI1105w	Phosphoglycerate kinase	5	▼
PfPk2	PFL1885c	Calcium dependent kinase	6	▼
PfTbp	PFL2345c	TATA binding homologue	7	▼
PCNA	PFL1285c	Proliferating cell nuclear antigen	8	▼
PP1-like	PF14_0224	Phosphatase	9	▼
Histone H3	MAL6P1.248	Histone	10	▼
Histone H2A	MAL6P1.249	Histone	11	▼
PfPk5	MAL13P1.279	Cyclin dependent kinase	1	▲

[a] Accession number according to PlasmoDB annotation.
[b] Genes are numbered as in Figure 4.
[c] Gene exhibiting an altered expression microarray experiments; ▲ stands for up-regulated and ▼ stands for down-regulated.

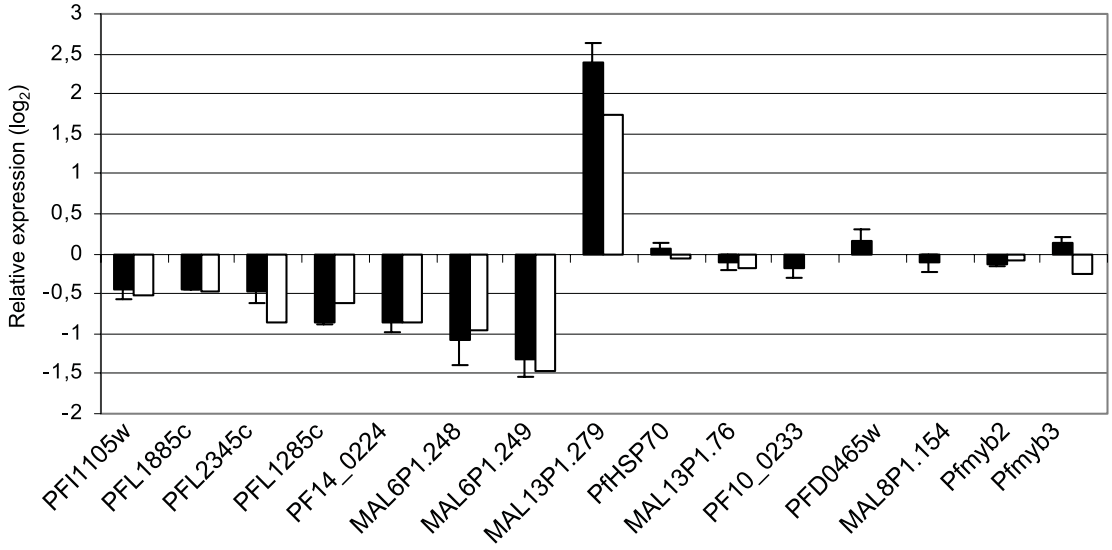


Figure 5. Comparison of microarray and RT-qPCR data. Microarray data are indicated by open bars. RT-qPCR data are indicated by filled bars. Relative transcriptional level of each gene (accession number indicated on bottom) is expressed as a $\log_2$ ratio of *pfmyb1* dsRNA-treated samples over untreated samples.

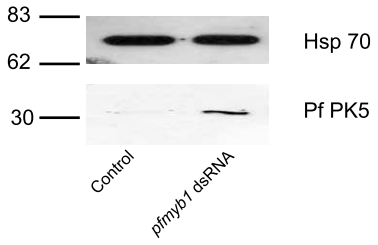


Figure 6. PfPk5 expression in *pfmyb1* dsRNA-treated samples. Trophozoite protein extracts were subjected to SDS-10% PAGE followed by Western blotting with either anti-PfHSP70 serum (upper panel) or anti-PfPk5 serum (lower panel) as indicated on the right side of the Figure. The protein markers are indicated on the left side of the Figure. Samples of untreated culture (control) and *pfmyb1* dsRNA-treated culture (*pfmyb1* dsRNA) were investigated. The Figure presents one representative experiment among three.

After purification of the DNA fragments, PCR analysis was carried out for the eight previously selected promoters. In Figure 7, odd numbers correspond to PCR performed on input DNA, and even numbers to immuno-precipitated DNA.

Promoters of three genes containing MRE (PF08_0054, PF10_0159, PF11_0096), but whose expression was not altered in parasites cultured in the presence of *pfmyb1* dsRNA in microarray experiments, were amplified in input (Figure 7 lanes 17, 19 and 21) but not ChIP samples (Figure 7 lanes 18, 20 and 22), even though expression of their cognate transcript peaked at the trophozoite stage.[26] In addition, the promoter of the PF07_0073 gene, devoid of MRE, yielded identical results (Figure 7 lanes 15–16).

In contrast, among the eight genes whose expression was altered in the presence of *pfmyb1* dsRNA, six (MAL13P1_279, PFI1105w, PFL1285c, MAL6P1.248, MAL6P1.249 and PF14_0224) gave rise to amplification products with both anti-PfMyb1 immuno-precipitated (Figure 7 lanes 2, 4, 10, 12 and 14) and input DNA (Figure 7 lanes 1, 3, 9,

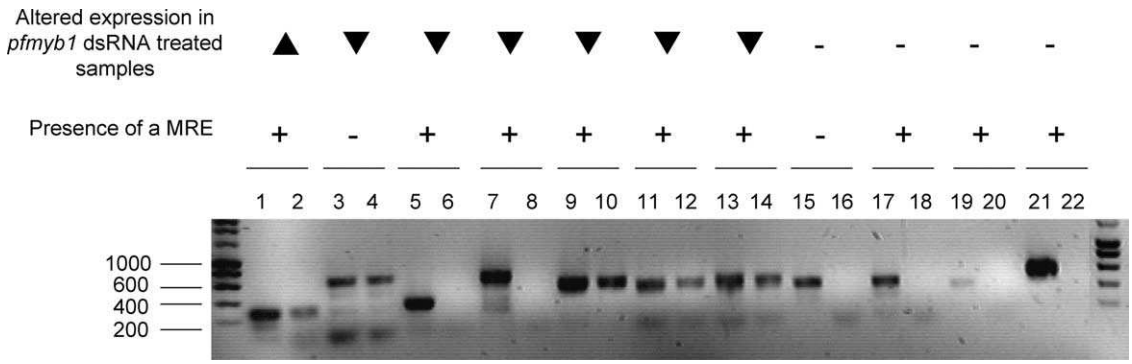


Figure 7. Chromatin immuno-precipitation (ChIP) using an anti-PfMyb1 serum. After DNA/PfMyb1 cross-linking within the parasite nuclei and immuno-precipitation of DNA using anti-PfMyb1 antibody, PCR was carried out with input DNA odd numbers) and immuno-precipitated DNA (even numbers) using primers amplifying promoters of the following genes: MAL13P1_279 (lane 1 and 2), PFI1105w (lane 3 and 4), PFL1885c (lane 5 and 6), PFL2345c (lane 7 and 8), PFL1285c (lane 9 and 10), MAL6P1.248 and MAL6P1.249 (lane 11 and 12), PF14_084 (lane 13 and 14), PF07_073 (lane 15 and 16), PF11_0096 (lane 17 and 18), PF08_0054 (lane 19 and 20) and PF10_0159 (lane 21 and 22). Amplicons were fractionated on 1.2% agarose gel. Genes identified in previous experiments as displaying up-regulated (▲), down-regulated (▼) or unaltered expression (–) are indicated. The presence (+) or absence (−) of Myb-related elements (MRE) in the promoter of the gene is indicated. Figure 7 portrays one representative experiment among three.

11 and 13). These six genes corresponded to five promoters, since the genes encoding histones H3 and H2A (MAL6P1.248 and MAL6P1.249) share a bidirectional promoter. Finally, the promoters of two (PFL1885c and PFL2345c) of these eight genes gave no amplification products using ChIP (Figure 7 lanes 6 and 8), in contrast to input samples (Figure 7 lanes 5 and 7).

Discussion

In light of the crucial role of transcriptional regulation during the complex life cycle of *P. falciparum*, we decided to investigate PfMyb1, a protein belonging to the tryptophan cluster family. Indeed, previous experiments have shown that this protein interacts specifically with several MRE.[22] These data, together with computational study of the PfMyb1 amino acid sequence,[22] have led us to consider this protein to be a genuine transcription factor.

In order to investigate the role of PfMyb1 in transcriptional regulation of *P. falciparum*, we decided to interfere, by means of long *pfmyb1* dsRNA, with its transcript and therefore its protein expression at the time of its highest expression, i.e. during the trophozoite stage.[22,26,27] As shown in Figure 1(a), a 40% growth inhibition was observed when parasite cultures were treated with *pfmyb1* dsRNA as compared with either untreated parasite cultures or cultures treated with unrelated long dsRNA. The observed inhibition thus appeared to depend specifically on *pfmyb1* dsRNA. This observation was made by using three different methods and confirmed by repeated experiments that consistently yielded similar results. The unrelated dsRNA was chosen for its relatively high AT-content (60%) and its length (570 nt), approximating those of the *pfmyb1* dsRNA (70% AT and 634 nt, respectively). Moreover, using the BLASTN tool of PlasmoDB, we did not find any detectable homology between this unrelated sequence and the entire *P. falciparum* genome. We further showed that parasite death occurred during the trophozoite to schizont transition, notably after the *pfmyb1* expression peak (Figure 1(b)).

Since addition of *pfmyb1* dsRNA modified parasite growth and development, we investigated the level of the *pfmyb1* transcript at the trophozoite stage. By two different experimental approaches based on RT-PCR (Figure 2) and performed with different sets of primers, we showed that *pfmyb1* dsRNA treatment diminished the *pfmyb1* transcript in the parasite by 75%. Moreover, the decrease in the *pfmyb1* transcript was correlated to a decrease in the corresponding protein, as determined by serial immuno-precipitation and Western blotting of the trophozoite protein extracts, using anti-PfMyb1 antibodies[22] and after normalization for protein loading. In conclusion, the decreased availability of *pfmyb1* transcript and protein, together with the growth inhibition occurring during the trophozoite to schizont transition, argue in favour of a major function for this transcription factor during the parasite erythrocytic cycle. This assumption is strengthened by the high level of conservation of *pfmyb1* gene within the diverse *Plasmodium* species listed in PlasmoDB.

In *Plasmodium*, the mechanisms by which gene expression is modified by dsRNA are still not clearly defined. The characterization of such mechanisms, whether involving classical RNA interference as suggested by previous work,[28,29] or an anti-sense effect as shown by Noonpakdee *et al.*,[30] is beyond the scope of this study. Herein, we took advantage of long dsRNA to reduce the cognate messenger and encoded protein, so as to study the consequences of this depletion on gene expression. Interestingly, many antisense transcripts have been

shown to arise during the erythrocytic cycle[31,32] and have been suggested to be potential regulatory elements in gene transcription.[33] These studies might imply that the machinery required to process these antisense transcripts and the resulting dsRNA is present in *P. falciparum*.

In order to identify at least some of the transcripts whose expression is governed by the transcription factor, we compared the transcriptome of cultures treated with *pfmyb1* dsRNA and control parasite cultures by means of our thematic DNA microarray. The slides were hybridized with cDNA prepared from total RNA of control and *pfmyb1*-treated cultures radiolabelled as reported.[24] Total RNA was extracted at the trophozoite stage prior to parasite death and when maximum expression of the *pfmyb1* transcript was reached (Figure 1). When unrelated dsRNA-treated samples were compared with untreated controls, no significant differential expression was identified by SAM (see Supplementary Data 2). The decrease in *pfmyb1* transcript and protein mediated by *pfmyb1* dsRNA altered the relative expression of a set of genes in four independent biological experiments normalized and filtered through SAM software (Figure 4).

In particular, among the 153 *P. falciparum* genes printed on the microarray, the level of expression of 128 transcripts was similar in control and in *pfmyb1* dsRNA-treated parasites (see Supplementary Data 1), suggesting that treatment with *pfmyb1* dsRNA did not markedly alter synchronization and development of the parasites. In this regard, upon examination of Giemsa-stained blood smears no substantial shift was detected between the two growth conditions.

In contrast, 11 genes were identified by SAM as differentially expressed in control and in *pfmyb1* dsRNA-treated parasites, of which one was up-regulated and ten down-regulated. All of the microarray data leading to the identification of the genes of interest as well as several control genes were verified by quantitative real-time RT-PCR. Three genes identified by SAM were eliminated upon performing independent biological experiments. It is noteworthy that these genes were very close to the limit defined by SAM (2, 3 and 4 of Figure 4). The remaining eight genes, listed in Table 1, gave similar expression ratios when microarray data were compared with real-time RT-PCR data. In addition, the expression ratio of seven other gene controls (Figure 5, right part) was verified, confirming the reliability of the microarray experiments. The level of expression of *pfmyb2* (PF10_0327) and *pfmyb3* (PF10_0143), two genes belonging to the tryptophan cluster family of transcription factors, was not modified by *pfmyb1* dsRNA. This was also the case for three genes showing slight but detectable homology to the *pfmyb1* DNA sequence used to generate dsRNA. Taken together, these data indicate that the activity of *pfmyb1* dsRNA is sequence-dependent.

Among the eight genes cited as differentially expressed in *pfmyb1* dsRNA-treated samples, the upregulation of PfPk5 (MAL13P1_279) expression observed at the transcript level was also detected at the protein level by Western blotting (Figure 6). It is worthy of note that PfPk5 has been annotated as a cdc2-related cyclin-dependent kinase bearing 60% identity to the human cdc2 gene.[34] In all eukaryotes, cyclin-dependent kinases are key regulators of cell cycle control and progression.[35] PfPk5 has been implicated in regulation of the nuclear division cycle[36] and is active during the erythrocytic cycle, and shares common features with the cdc-related kinase activation process.[37] Taken together, these data suggest a crucial role for PfPk5 in regulating cell cycle progression of *P. falciparum*. A clear link between the classical cell cycle (G1, S, G2 and M phases) and the *Plasmodium* erythrocytic cell cycle has not been established.[38] Nevertheless, DNA synthesis appears to occur during the trophozoite stage and asynchronous nuclear division during the trophozoite to schizont transition.[39] These two critical phenomena, however, may overlap during trophozoite to schizont transition. Indeed, PfPk5 may play a crucial role during the trophozoite to schizont transition and its over-expression in *pfmyb1* dsRNA-treated culture may explain the growth inhibition and parasite death at that particular time in the erythrocytic cycle.

Furthermore, in *pfmyb1* dsRNA-treated samples, a decreased expression in the gene PFL1285c, a homologue of proliferating cell nuclear antigen (PCNA), was observed. In eukaryotes, PCNA has been shown to be a key player in DNA replication, DNA repair and cell cycle control.[40] In *Plasmodium*, PCNA1 expression was maximum at the trophozoite stage at the time of DNA replication[41] and its altered expression could therefore have played a role in the death of *pfmyb1* dsRNA-treated parasites. Alteration of the glucose metabolism *via* the partial inhibition of expression of the phosphoglycerate kinase (PGK, PFI1105w) might also have participated in the alteration of the parasite life cycle. PGK has been found to be a key enzyme of the glucose pathway and its activity has been detected during the erythrocytic cycle.[42] Of course, all of the observed variations in gene expression might contribute in a cooperative manner to the alteration of parasite development and therefore participate in promoting parasite death. Finally, two histone (H3 and H2A) genes have been shown to be under-expressed in *pfmyb1* dsRNA-treated samples, in contrast to the two other histones (H4 and H2B). The two histone (H3 and H2A) genes appeared to be under the control of a bidirectional promoter. This observation provided another argument in favour of an activity of PfMyb1 as a transcription factor.

In order to determine, among the eight genes identified by microarray, those whose expression might be directly controlled by PfMyb1, we performed a chromatin immuno-precipitation assay with a specific anti-PfMyb1 antibody, followed by a PCR identification of gene promoters. For six genes (MAL13P1_279, PFI1105w, PFL1285c,

MAL6P1.248, MAL6P1.249 and PF14_0224), an interaction was observed *ex vivo* between the promoter and the transcription factor PfMyb1, suggesting that they might be directly regulated by PfMyb1. By contrast, two genes (PFL1885c and PFL2345c, see Figure 7, lanes 5–8) were not amplified in ChIP samples. These genes had in fact shown small variation in the level of gene expression in *pfmyb1* dsRNA samples (spots 5 and 6 of Figure 4). The absence of amplification might be attributable to a weak interaction between their potential MRE and PfMyb1, below the threshold of detection of ChIP methodology, or to poor accessibility of the MRE within the chromatin. Alternatively, these genes might be regulated by a transcription factor itself under PfMyb1 control. Similar results were obtained with three control genes (PF08_0054, PF10_0159, PF11_0096) displaying no significant differential expression in microarray experiments and sharing the same peak of expression during the trophozoite stage. Identical results were obtained as regards the PF07_0073 promoter that lacks MRE (Figure 7).

Finally, the PFI1105w promoter, despite the apparent absence of MRE, gave rise to a positive signal in ChIP experiments (Figure 7, lanes 3 and 4). This result has been shown to arise with a Myb protein from *Saccharomyces cerevisiae* incapable of binding conventional MRE.[43] *P. falciparum* might have evolved new PfMyb1/DNA binding specificities that might depend on chromatin context and/or availability of the protein partner. Indeed, identification of binding sites for PfMyb1 in the PGK promoter is a new challenge for understanding PfMyb1-dependent transcriptional regulation. Nevertheless, this experiment clearly demonstrated the ability of the PfMyb1 protein to bind to the promoter of a set of genes of critical importance in the *P. falciparum* cell cycle and therefore to regulate their expression in a direct manner. Moreover, these data suggest that PfMyb1 may participate directly in either repression (for expression of PfPk5) or activation (for the five other genes) of transcriptional gene expression. This is a common feature for Myb proteins, among which is the c-Myb factor reported to mediate either up or down-regulation of different genes.[20] Hence, depending on post-translational modifications of the transcription factors,[44,45] the protein partners present on each given promoter and the chromatin context, the activity of transcription factors could be modulated either for activation or repression of transcription.

In conclusion, we describe the first attempt to unravel the PfMyb1 regulation network. Although more work is needed to understand the multiple determinants of PfMyb1-mediated regulation, these experiments provide important clues for future analysis of factors determining regulation of gene expression by PfMyb1. The data reported herein suggest that PfMyb1 is an essential transcription factor for the erythrocytic cycle of *P. falciparum* and, moreover, that PfMyb1 directly regulates key genes involved in cell cycle regulation and progression.

Materials and Methods

Plasmodium falciparum culture

The 3D7 clone of *P. falciparum* was provided by Dr D. Walliker and was cultured as described[46] with slight modifications. To obtain synchronized erythrocytic stages, ring cultures were enriched by three successive treatments with 5% (w/v) sorbitol at 44 hour intervals.[47] Synchronization of culture was verified by morphological analysis of Giemsa-stained blood smears at different time-points.

dsRNA preparation

A polymerase chain reaction (PCR) fragment (634 nt) representing the *pfmyb1* gene (PF13_0088) was amplified from 3D7 genomic DNA using forward (5′GGATG AAATGTGATGATAAAAC) and backward primers (5′TACTTCATCTTTTGTCCATTTT). The PCR product was cloned into TOPO PCR II vector (Invitrogen). A PCR fragment (570 nt) representing the portion of the *env* gene encoding gp120 (GeneBank accession number K02013) was amplified from HIV-1 genomic DNA using forward (5′ATGCCCCAGACTGTGAGTTG) and backward primers (5′CTACTGTAATTCAACACAACT) and cloned into pBluescript SK− vector. The clones were sequenced by dideoxy sequencing reactions. Plasmids were used as a template to generate sense RNA (sRNA) and antisense RNA (asRNA) using T7, SP6 and T3 RiboMAX Express Large Scale RNA Production System (Promega). The dsRNA was prepared as follows: equal amounts of sRNA and asRNA were separately denatured at 95 °C for one minute, mixed and then heated to 95 °C for an additional minute. Annealing was performed by slow cooling over several hours and thereafter dsRNA samples were analyzed on native 1% (w/v) agarose gel.

dsRNA assay

For each experiment using dsRNA, 25 μg of dsRNA was added per millilitre of highly synchronized young ring culture (four to six hours post invasion). Parasitaemia was set as described in the next section for manual counting, [^{3}H]hypoxanthine uptake assay and flow cytometry in 24-well plates and incubated for 48 hours. A 20 ml culture of 5% parasitaemia of highly synchronized young ring stages was set for RNA and protein extraction. This culture was incubated in 75 cm^3 culture flask for 18–20 hours as described above *P. falciparum* culture. Control cultures were carried out as described above and an equivalent volume of dsRNA annealing buffer was added to the medium.

Parasite growth assay

[^{3}H]hypoxanthine uptake was evaluated as described[48] with slight modifications. In brief, 200 μl of ring stage parasitized erythrocytes (parasitaemia, 0.5%; hematocrit, 1.8%) were dispensed in 96-well plates preloaded with 25 μg of dsRNA per millilitre in triplicate and with serial dilutions of chloroquine in positive control wells. After 24 hours, [^{3}H]hypoxanthine was added to each well and plates were then incubated for an

additional 24 hours. Parasites were harvested, and incorporation of radioactivity was determined by liquid scintillation counting. Experiments were repeated twice. Data are reported as mean and standard deviation (SD) of percentage parasite growth inhibition in triplicate experiments relative to untreated controls after correcting for background [^{3}H]hypoxanthine incorporation in uninfected erythrocytes. Alternatively, growth inhibition was determined by microscopic examination of Giemsa-stained thin blood smears. Smears from untreated cultures were always used as controls. Parasitaemia was measured by counting 3000 red blood cells and expressed as percentage of total parasitized erythrocytes. Twenty-five micrograms of dsRNA was added to a ring stage culture (1.5% parasitaemia, 5% hematocrit) and parasitaemia was evaluated 48 hours later. Experiments were repeated nine times in triplicate. Finally, flow cytometry was performed on 48 hour-treated culture (1.5% rings, 5% hematocrit) as described.[49] Experiments were repeated three times in triplicate. Statistical analysis was performed using Student's *t*-test.

RT-PCR and RT-qPCR

RT-PCR: total RNA (1 μg) was treated with two units of RNase-free DNase I (Qiagen) and purified using RNeasy columns with the Qiagen clean-up procedure to avoid DNA contamination. Synthesis of cDNA was performed for 50 minutes at 42 °C using 1 μg of RNA (in 40 μl final volume), 50 units of MMLV reverse transcriptase (Superscript II Invitrogen), 0.5 ng of random hexamers, 5 mM $MgCl_2$, 20 units of RNaseout (Invitrogen), 10 mM DTT and 0.25 mM each dNTP. Samples were then subjected to RNase-H treatment with two units of enzyme for 20 minutes at 37 °C. PCR was carried out with 2 μl of each cDNA sample (or the negative control) in the presence of 1 μM primers; 18 S forward (5′GACTCAACACGGGGAAACTCACTAGTT) and backward primers (5′ACAATTCATCATATCTTTAATCG GTA) or *pfmyb1* forward (5′GGATGAAATGTGATGA TAAAAC) and backward primers (5′TACTTCATC TTTTGTCCATTTT or 5′TCTCCTTTTGTAAGATATTT CATCATTCA) or MAL13P1.76 forward (5′ATGCAA AATCCTCAAAAT) and backward (5′TTATGTATCGT TAATTAAAC) primers, 200 μM dNTP, and two units of Taq polymerase enzyme. After denaturation at 95 °C for two minutes, 30 cycles with annealing at 55 °C, elongation at 60 °C and denaturation at 95 °C, each step for one minute, were performed to amplify the *pfmyb1* and MAL13P1.76 fragments. Twenty cycles of the same amplification profile were used to amplify the ribosomal 18 S subunit. We verified that, under these conditions and number of cycles, the magnitude of the signals remained proportional to RNA concentrations, making the RT-PCR a semi-quantitative procedure. Fragments of expected lengths (657 bp, 1204 bp and 634 bp or 1015 bp for 18 S, MAL13P1.76 and *pfmyb1*, respectively) were observed by 1% agarose gel electrophoresis and relative amounts of mRNA were determined by densitometry (Densylab, Microvision Instruments, Evry, France).

Real-time quantitative PCR reaction was performed as described.[24] Specific primers are listed in Supplementary Data 1.

Multiplex real-time quantitative PCR was performed with LUX-primers (Invitrogen). Reactions were performed in a 25 μl volume containing 5 μl of 25-fold diluted cDNA preparations, 12.5 μl of Platinum PCR mix (Invitrogen) supplemented with 3 mM $MgCl_2$, and 1875 nM *pfmyb1* primers and 94 nM 18 S primers. Amplification and detection of specific products were performed with the MX4000 light cycler (Stratagene) with the following cycle profile: one cycle at 50 °C for two minutes, one cycle at 95 °C for two minutes, 40 cycles with 15 seconds denaturation at 95 °C and 30 seconds annealing-elongation at 60 °C. The 18 S primers were CACCAT TTTCTTGATTTCTTGGATGG labelled with JOE, and GATCTCGTTCGTTATCGGAAT. *pfmyb1* primers were CACATCTGCAAAGGAATGTCAAAAGATG labelled with FAM, and CACGACAAAGGACTTCTATTGGT.

The size of each PCR product and number of bands were verified on a 10% (w/v) acrylamide gel. Two additional reactions were performed, either without reverse transcriptase or without the RNA sample, to verify the absence of DNA contamination. The quantity of cDNA for each experimental gene was normalized to the 18 S concentration (chr5.rRNA-1-18s-A) in each sample. For each sample and each gene, experiments were performed twice and in triplicate. Relative gene expression was expressed as the $\log_2$ of the ratio of expression at each time-point over that of the ring stage using the $2^{-\Delta\Delta CT}$ method.

Serial immuno-precipitation and Western blotting

Protein extracts of *P. falciparum* trophozoites were prepared as described.[16] Protein (100 μg) was subjected to immuno-precipitation using 2×10^7 magnetic tosylactivated Dynabeads (M-280, Dynal Biotech) coated with 2.5 μg of either anti-PfMyb1 or preimmune serum, according to the manufacturer's recommendations. Immuno-precipitated proteins were resolved by SDS-10% PAGE and subjected, after transfer onto nitrocellulose membrane, to Western blotting using the anti-PfMyb1. The membrane was first incubated overnight at 4 °C with a 1:1000 dilution of preimmune or anti-PfMyb1 serum and then incubated with a 1:10,000 dilution of peroxidase-conjugated anti-rabbit antibody (Biolabs). Concomitantly, 5 μg of the supernatant proteins was resolved by SDS-10% PAGE and blotted onto nitrocellulose membrane. The membrane was first incubated overnight at 4 °C with a 1:10,000 dilution of anti-PfHsp70 serum and then with a 1:10,000 dilution of peroxidase-conjugated anti-mouse antibody (Biolabs).

Western blotting experiments on PfPk5 protein were performed on 60 μg of protein extracts prepared as described.[50] Proteins were resolved by SDS-12% PAGE and subjected, after transfer onto nitrocellulose membrane, to Western blotting using the anti-PfPk5 serum, a kind gift from C. Doerig. The membrane was first incubated overnight at 4 °C with a 1:1000 dilution anti-PfPk5 serum and then with a 1:10,000 dilution of peroxidase-conjugated anti-rabbit antibody (Biolabs). PfHSP70 protein was detected as described above.

All membranes were developed using a chemiluminescent substrate, according to the manufacturer's instructions (ECL, Amersham-Pharmacia Biotech). Relative amounts of proteins were determined by densitometry (Densylab, Microvision Instruments, Evry, France).

Microarray construction, probe labeling, hybridization and data analysis

Microarrays were prepared as described by Gissot *et al*.[24] Briefly, 17 spots representing positive and negative controls such as sequences unrelated to *P. falciparum* were printed. Alien PCR products from Stratagene, with no apparent similarity to the *P. falciparum* genomic sequence,

were used for normalization of the reverse transcription. 3D7 genomic DNA was also printed in order to normalize the total amount of each cDNA present during the hybridization. All targets were spotted in duplicate with a Qarray arrayer (Genetix) onto poly-L-lysine-coated slides prepared as described[51] with a spot spacing of 400 μm, centre to centre onto a 1 cm^2 area.

Probe labeling and hybridization were performed as described.[24] Data were collected as an image file, gridded, and converted into an Excel file using β-Vision software (Biospace mesure). The global background was subtracted for each signal. In addition, the signal intensity of each spot was normalized according to the average value of intensities of spots corresponding to exogenous control PCR products (Alien, Stratagene) on each array in order to control for variation in the efficiency of reverse transcription. Furthermore, *P. falciparum* genomic DNA was used to normalize the amount of cDNA present in the hybridization. Four independent biological experiments were performed. Statistical discrimination of genes expressed differentially in the two stages was performed on the complete normalized set of data using the freely available SAM software†.[25]

Chromatin immuno-precipitation

Potential MRE were localized in promoter sequences by using MatInspector[52] according to three matrix families (one from the plant and two from the vertebrate subsections). This protocol is adapted from Hecht *et al*.[53] A 100 ml sample of trophozoite culture was subjected to protein-DNA crosslink using 1% (v/v) formaldehyde (Sigma) at 37 °C for 15 minutes. Nuclei were prepared as described[16] and resuspended in RIPA buffer (30 mM Tris–HCl (pH 8), 150 mM NaCl, 20 mM $MgCl_2$, 1 mM EDTA, 1 mM DTT, 0.5% (v/v) Triton X-100, 1% (v/v) Nonidet-P40, 1 mM PMSF, 1X Complete TM mixture tablet from Roche Molecular Biochemicals). Samples were sonicated (BioBlock Scientific, VibraCell 72405) three times at 30 W for ten seconds, allowing 30 seconds cooling between sonifications, and centrifuged for five minutes at 14,400 ***g***. The supernatant, containing soluble chromatin, was incubated overnight at 4 °C with 2×10^7 magnetic tosylactivated Dynabeads (M-280, Dynal Biotech) coated with 2.5 μg of either preimmune or anti-PfMyb1 serum. Beads were washed twice with 1 ml of RIPA buffer supplemented with 0.01% (w/v) SDS. Immunoprecipitated chromatin was eluted using TE (10 mM Tris (pH 8), 1 mM EDTA) supplemented with 1% SDS. Chromatin was reverse cross-linked during six hours at 65 °C and subjected to proteinase K (20 μg) treatment for two hours at 37 °C. DNA was extracted with phenol/chloroform/isoamyl alcohol and precipitated with ethanol and sodium acetate. DNA from input and ChIP samples were resuspended in 200 μl and 20 μl of TE, respectively.

The amount of template used for PCR was 2 μl of a 1:100 dilution of input DNA and 2 μl of undiluted ChIP DNA. PCR was performed in the presence of 1 μM specific primers (see Supplementary Data 2), 200 μM dNTP, and two units of Taq polymerase. After denaturation at 95 °C for two minutes, 40 cycles were performed with annealing at 50 °C, elongation at 60 °C and denaturation at 95 °C, each step for one minute. PCR products were analyzed on 1.2% agarose gel.

† http://www-stat.stanford.edu/~tibs/SAM/stat.SAM/

Acknowledgements

We thank Dr Christian Doerig for the kind gift of the anti-PfPk5 serum and Dr Jennifer Richardson for critical reading of the manuscript. We are grateful to Professor Dominique Mazier (INSERM U511) for continuous encouragement and support. M.G., P.R. and C.B. were financially supported by the Ministère de l'Education Nationale, France. M.G. held a fellowship from Fondation pour la Recherche Médicale (FRM). The project was supported by INSERM funds to C.V. We are grateful to the Pitié-Salpêtrière genomic core facility (P3S) supported by INSERM, the région Ile- de-France, the Fondation de Recherche HMR-Aventis, and by University Paris VI.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2004.11.045

References

1. World Health Organization (2000). *WHO Expert Committee on Malaria—Twentieth Report*. WHO, Rome.
2. Demidov, V. V. (2003). Infection stage clues to new antimalarial medicines. *Drug Discov. Today*, **8**, 913–915.
3. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V. *et al*. (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science*, **282**, 1126–1132.
4. Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M. *et al*. (1999). The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature*, **400**, 532–538.
5. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W. *et al*. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.
6. Horrocks, P., Dechering, K. & Lanzer, M. (1998). Control of gene expression in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **95**, 171–181.
7. Levitt, A. (1993). RNA processing in malarial parasites. *Parasitol. Today*, **9**, 465–468.
8. Cary, C., Lamont, D., Dalton, J. P. & Doerig, C. (1994). *Plasmodium falciparum* chromatin: nucleosomal organisation and histone-like proteins. *Parasitol Res.* **80**, 255–258.
9. Li, W. B., Bzik, D. J., Gu, H. M., Tanaka, M., Fox, B. A. & Inselburg, J. (1989). An enlarged largest subunit of *Plasmodium falciparum* RNA polymerase II defines conserved and variable RNA polymerase domains. *Nucl. Acids Res.* **17**, 9621–9636.
10. Li, W. B., Bzik, D. J., Tanaka, M., Gu, H. M., Fox, B. A. & Inselburg, J. (1991). Characterization of the gene encoding the largest subunit of *Plasmodium falciparum* RNA polymerase III. *Mol. Biochem. Parasitol.* **46**, 229–239.
11. Fox, B. A., Li, W. B., Tanaka, M., Inselburg, J. & Bzik, D. J. (1993). Molecular characterization of the largest subunit of *Plasmodium falciparum* RNA polymerase I. *Mol. Biochem. Parasitol.* **61**, 37–48.

12. McAndrew, M. B., Read, M., Sims, P. F. & Hyde, J. E. (1993). Characterisation of the gene encoding an unusually divergent TATA-binding protein (TBP) from the extremely A+T-rich human malaria parasite *Plasmodium falciparum*. *Gene*, **124**, 165–171.
13. Hirtzlin, J., Farber, P. M. & Franklin, R. M. (1994). Isolation of a novel *Plasmodium falciparum* gene encoding a protein homologous to the Tat-binding protein family. *Eur. J. Biochem.* **226**, 673–680.
14. Horrocks, P. & Kilbey, B. J. (1996). Physical and functional mapping of the transcriptional start sites of *Plasmodium falciparum* proliferating cell nuclear antigen. *Mol. Biochem. Parasitol.* **82**, 207–215.
15. Lanzer, M., de Bruin, D. & Ravetch, J. V. (1992). Transcription mapping of a 100 kb locus of *Plasmodium falciparum* identifies an intergenic region in which transcription terminates and reinitiates. *EMBO J.* **11**, 1949–1955.
16. Osta, M., Gannoun-Zaki, L., Bonnefoy, S., Roy, C. & Vial, H. J. (2002). A 24 bp *cis*-acting element essential for the transcriptional activity of *Plasmodium falciparum* CDPdiacylglycerol synthase gene promoter. *Mol. Biochem. Parasitol.* **121**, 87–98.
17. Miller, L. H., Baruch, D. I., Marsh, K. & Doumbo, O. K. (2002). The pathogenic basis of malaria. *Nature*, **415**, 673–679.
18. Calderwood, M. S., Gannoun-Zaki, L., Wellems, T. E. & Deitsch, K. W. (2003). *Plasmodium falciparum* var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron. *J. Biol. Chem.* **278**, 34125–34132.
19. Dechering, K. J., Kaan, A. M., Mbacham, W., Wirth, D. F., Eling, W., Konings, R. N. & Stunnenberg, H. G. (1999). Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell. Biol.* **19**, 967–978.
20. Kanei-Ishii, C., Nomura, T., Ogata, K., Sarai, A., Yasukawa, T., Tashiro, S. *et al*. (1996). Structure and function of the proteins encoded by the *myb* gene family. *Curr. Top Microbiol. Immunol.* **211**, 89–98.
21. Lipsick, J. S. (1996). One billion years of Myb. *Oncogene*, **13**, 223–235.
22. Boschet, C., Gissot, M., Briquet, S., Hamid, Z., Claudel-Renard, C. & Vaquero, C. (2004). Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 *Plasmodium falciparum* clones. *Mol. Biochem. Parasitol.* **138**, 159–163.
23. Ness, S. A., Marknell, A. & Graf, T. (1989). The v-myb oncogene product binds to and activates the promyelocyte-specific mim-1 gene. *Cell*, **59**, 1115–1125.
24. Gissot, M., Refour, P., Briquet, S., Boschet, C., Coupe, S., Mazier, D. & Vaquero, C. (2004). Transcriptome of 3D7 and its gametocyte-less derivative F12 *Plasmodium falciparum* clones during erythrocytic development using a gene-specific microarray assigned to gene regulation, cell cycle and transcription factors. *Gene*, **341**, 267–277.
25. Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
26. Bozdech, Z., Llinas, M., Pulliam, B. L., Wong, E. D., Zhu, J. & DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1**, 5.
27. le Roch, K. G., Johnson, J. R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M. *et al*. (2004). Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res.* **14**, 2308–2318.
28. Malhotra, P., Dasaradhi, P. V., Kumar, A., Mohmmed, A., Agrawal, N., Bhatnagar, R. K. & Chauhan, V. S. (2002). Double-stranded RNA-mediated gene silencing of cysteine proteases (falcipain-1 and -2) of *Plasmodium falciparum*. *Mol. Microbiol.* **45**, 1245–1254.
29. McRobert, L. & McConkey, G. A. (2002). RNA interference (RNAi) inhibits growth of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **119**, 273–278.
30. Noonpakdee, W., Pothikasikorn, J., Nimitsantiwong, W. & Wilairat, P. (2003). Inhibition of *Plasmodium falciparum* proliferation *in vitro* by antisense oligodeoxynucleotides against malarial topoisomerase II. *Biochem. Biophys. Res. Commun.* **302**, 659–664.
31. Gunasekera, A. M., Patankar, S., Schug, J., Eisen, G., Kissinger, J., Roos, D. & Wirth, D. F. (2004). Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol.* **136**, 35–42.
32. Patankar, S., Munasinghe, A., Shoaibi, A., Cummings, L. M. & Wirth, D. F. (2001). Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. *Mol. Biol. Cell.* **12**, 3114–3125.
33. Kyes, S., Christodoulou, Z., Pinches, R. & Newbold, C. (2002). Stage-specific merozoite surface protein 2 antisense transcripts in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **123**, 79–83.
34. Ross-Macdonald, P. B., Graeser, R., Kappes, B., Franklin, R. & Williamson, D. H. (1994). Isolation and expression of a gene specifying a cdc2-like protein kinase from the human malaria parasite *Plasmodium falciparum*. *Eur. J. Biochem.* **220**, 693–701.
35. Morgan, D. O. (1997). Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu. Rev. Cell. Dev. Biol.* **13**, 261–291.
36. Graeser, R., Wernli, B., Franklin, R. M. & Kappes, B. (1996). *Plasmodium falciparum* protein kinase 5 and the malarial nuclear division cycles. *Mol. Biochem. Parasitol.* **82**, 37–49.
37. Graeser, R., Franklin, R. M. & Kappes, B. (1996). Mechanisms of activation of the cdc2-related kinase PfPK5 from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **79**, 125–127.
38. Arnot, D. E. & Gull, K. (1998). The Plasmodium cell-cycle: facts and questions. *Ann. Trop. Med. Parasitol.* **92**, 361–365.
39. Read, M., Sherwin, T., Holloway, S. P., Gull, K. & Hyde, J. E. (1993). Microtubular organization visualized by immunofluorescence microscopy during erythrocytic schizogony in *Plasmodium falciparum* and investigation of post-translational modifications of parasite tubulin. *Parasitology*, **106**, 223–232.
40. Kelman, Z. (1997). PCNA: structure, functions and interactions. *Oncogene*, **14**, 629–640.
41. Kilbey, B. J., Fraser, I., McAleese, S., Goman, M. & Ridley, R. G. (1993). Molecular characterisation and stage-specific expression of proliferating cell nuclear antigen (PCNA) from the malarial parasite, *Plasmodium falciparum*. *Nucl. Acids Res.* **21**, 239–243.
42. Grall, M., Srivastava, I. K., Schmidt, M., Garcia, A. M., Mauel, J. & Perrin, L. H. (1992). *Plasmodium falciparum*: identification and purification of the phosphoglycerate kinase of the malaria parasite. *Expt. Parasitol.* **75**, 10–18.

43. Hovring, I., Bostad, A., Ording, E., Myrset, A. H. & Gabrielsen, O. S. (1994). DNA-binding domain and recognition sequence of the yeast BAS1 protein, a divergent member of the Myb family of transcription factors. *J. Biol. Chem.* **269**, 17663–17669.
44. Bannister, A. J. & Miska, E. A. (2000). Regulation of gene expression by transcription factor acetylation. *Cell. Mol. Life Sci.* **57**, 1184–1192.
45. Whitmarsh, A. J. & Davis, R. J. (2000). Regulation of transcription factor function by phosphorylation. *Cell. Mol. Life Sci.* **57**, 1172–1183.
46. Trager, W. & Jensen, J. B. (1976). Human malaria parasites in continuous culture. *Science*, **193**, 673–675.
47. Lambros, C. & Vanderberg, J. P. (1979). Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. *J. Parasitol.* **65**, 418–420.
48. Desjardins, R. E., Canfield, C. J., Haynes, J. D. & Chulay, J. D. (1979). Quantitative assessment of antimalarial activity *in vitro* by a semiautomated microdilution technique. *Antimicrob. Agents Chemother.* **16**, 710–718.
49. Makler, M. T., Lee, L. G. & Recktenwald, D. (1987). Thiazole orange: a new dye for Plasmodium species analysis. *Cytometry*, **8**, 568–570.
50. Merckx, A., le Roch, K., Nivez, M. P., Dorin, D., Alano, P., Guiterrez, G. J. *et al*. (2003). Identification and initial characterization of three novel cyclin-related proteins of the human malaria parasite *Plasmodium falciparum*. *J. Biol. Chem.* **278**, 39839–39850.
51. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
52. Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.* **23**, 4878–4884.
53. Hecht, A., Strahl-Bolsinger, S. & Grunstein, M. (1999). Mapping DNA interaction sites of chromosomal proteins. Crosslinking studies in yeast. *Methods Mol. Biol.* **119**, 469–479.

Edited by J. Karn

(Received 17 September 2004; received in revised form 18 November 2004; accepted 18 November 2004)

RESUME

Plasmodium falciparum est un parasite dont le développement érythrocytaire est composé de deux phases successives : une prolifération intense responsable de la maladie (phase asexuée) et une différenciation en gamétocytes responsable de la dissémination du parasite (phase sexuée). Ce changement de statut de la cellule serait dû en partie à une expression différentielle des gènes, notamment à une régulation au niveau de la transcription. Cette régulation nécessite l'interaction de deux protagonistes dont la caractérisation devrait aboutir à une meilleure connaissance du développement du parasite et permettre de trouver de nouvelles voies pour combattre la maladie.

i) Après identification des promoteurs des gènes, des éléments connus chez les autres eucaryotes ainsi que des éléments dits spécifiques de *P. falciparum* ont été recherchés à l'aide de différents programmes bioinformatiques, puis regroupés en modules. Les familles de gènes, dont l'expression est coordonnée ou altérée par l'expression diminuée d'un fateur de transcription, devraient partager au sein de leurs promoteurs des éléments de régulation leur permettant d'être exprimées à un moment précis du développement. L'identification d'éléments de régulation spécifiques du parasite peut conduire à l'identification de facteurs de transcription spécifiques, eux aussi, du parasite.

ii) Des facteurs impliqués dans la régulation transcriptionnelle, se liant à l'ADN de manière séquence-spécifique ou structure-spécifique, ont été recherchés dans le génome de *Plasmodium* par homologie de séquences. Des phases ouvertes de lecture codant pour des facteurs appartenant aux familles de protéines à domaine Myb, à doigt de zinc ou encore présentant architécture β ont été identifiées. Le clonage et la caractérisation biochimique de trois de ces facteurs ont confirmé la pertinence de la mise en évidence informatique de ces protéines. La présence de ces facteurs dans le parasite et leur disponibilité au cours du développement ont été déterminées ainsi que leur capacité à se fixer à l'ADN par l'intermédiaire d'une structure caractéristique de la chromatine ou de leur élément de régulation.

Ce tavail représente une première étape vers la compréhension de la régulation transcriptionnelle des événements clés du cycle érythrocytaire de *P. falciparum* et permet d'ouvrir une nouvelle voie pour la lutte contre le paludisme.

DISCIPLINE

Analyse des génomes et modélisation moléculaire

MOTS-CLES

paludisme, *Plasmodium falciparum*, régulation, transcription, bioinformatique, éléments de régulation, facteurs de transcription.

ADRESSE DU LABORATOIRE

INSERM U511 - Immunobiologie cellulaire et moléculaire des infections parasitaires
CHU Pitié-Salpêtrière - 91 boulevard de l'Hôpital - 75013 PARIS