Dynamique des hélitrons dans le genome d'Arabidopsis thaliana : développement de nouvelles stratégies d'analyse des éléments transposables

Sébastien Tempel

N° d'ordre : 3558

THÈSE

présentée

DEVANT L'UNIVERSITÉ DE RENNES 1

pour obtenir

le grade de : DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention Biologie

PAR

TEMPEL Sébastien

Équipe d'accueil 1 : projet Symbiose (IRISA, Rennes)
Équipe d'accueil 2 : Expression génétique et adaptation (Ecobio, Rennes)

École Doctorale : Vie-Agro-Santé
Composantes universitaires : IFSIC, SVE

TITRE DE LA THÈSE :

DYNAMIQUE DES HÉLITRONS DANS LE GENOME D'ARABIDOPSIS THALIANA : DÉVELOPPEMENT DE NOUVELLES STRATÉGIES D'ANALYSE DES ÉLÉMENTS TRANSPOSABLES

SOUTENUE LE 18 juin 2007 devant la commission d'examen

COMPOSITION DU JURY :

Yves	Bigot	Rapporteur
Marie-France	Sagot	Rapporteur
Pierre	Capy	Examinateur
Pierre	Rouzé	Examinateur
Abdelhak	El Amrani	Directeur de thèse
Jacques	Nicolas	Co-Directeur de thèse

Remerciements

Je tiens à remercier en premier ma femme Laure, pour m'avoir accompagné tout au long de ma thèse, pour avoir corrigé mes nombreuses fautes d'orthographes et aussi pour m'avoir poussé à me surpasser.

Je remercie ma famille et en particulier ma mère, ma tante et ma marraine. Elles ont toujours été à mes côtés, m'ont toujours soutenu dans mes décisions et aidé à financer mes études. Je remercie aussi la famille de Laure ; ils m'ont accueilli à bras ouverts.

J'offre mes remerciements les plus sincères à mes directeurs de thèse : Jacques Nicolas, Abdelhak El Amrani et Ivan Couée. Je remercie Jacques, responsable du projet Symbiose à l'IRISA, pour sa bonne humeur, ses conseils éclairés et justes. Pour avoir essayé de m'enseigner la rigueur scientifique, propre aux informaticiens. Je salue Abdel(hak), pour son enthousiasme pour mes travaux, son sens des relations humaines, et pour m'avoir choisi lors du stage de DEA qui est le tout début de cette thèse. Et je remercie Ivan, pour sa pondération et sa justesse face à mon enthousiasme et pour ses remarques toujours pertinentes.

Je rends hommage à l'équipe Symbiose, l'équipe de bioinformatique où s'est déroulée ma thèse. Ils m'ont parfaitement intégré dans leur équipe. Nous avons partagé ensemble des bons moments au sein même mais aussi en dehors des bureaux (les séminaires d'équipe de juin, par exemple). Je les remercie encore de m'avoir donné de bons conseils tout au long de ce doctorat et aidé à rédiger cette thèse. Je veux envoyer aussi un grand merci à tous les anciens membres de l'équipe qui me manquent et à tous les membres actuels de l'équipe qui me manqueront quand je partirai. Je remercie aussi les membres de l'UMR ECOBIO qui ont toujours répondu avec joie à toutes les questions que j'ai pu leur poser.

Je remercie aussi tous les membres de mon jury de thèse pour avoir accepté de corriger ma thèse et particulièrement Yves Bigot pour tous ses conseils pendant ma thèse et ce bon hamburger en Californie.

Je remercie enfin tous mes amis, et en particulier "les grumeaux" (Yannick et Nicolas), pour toutes les bonnes soirées passées de jeux de sociétés, de rôles et les nombreux restaurants ...

Table des matières

Glossaire

$\in$	**inclus dans**
$\forall$	**quel que soit**
$\exists$	**il existe**
$\cup$	**union**
/	**tel que**
AA	**A**cide **A**miné
ABF1	**A**RS-**b**inding **f**actor 1
Ac	**Ac**tivateur
ADC	**A**rginine **Dec**arboxylase
ADN	**A**cide **D**ésoxyribo**N**ucléique
Amt	**A**SA1 **mut**ant
ARN	**A**cide **R**ibo**N**ucléique
ARNm	**ARN m**essager
ARNt	**ARN** de **t**ransfert
AVL	**A**lgorithme de **V**raisemblance du **L**ien
BAC	**B**acterial **A**rtificial **C**hromosome
CAH	**C**lassification **A**scendante **H**iérarchique
CHAVL	**C**lassification **H**iérarchique par **A**nalyse de la **V**raisemblance des **L**iens
ci	**c**omplémentaire-**i**nversé
DAWG	**D**irected **A**cyclic **W**ord **G**raph
Ds	**Dis**sociation
ET	**E**lément **T**ransposable
gag	**g**roup **a**ssociated **a**ntigen
HMM	**H**idden **M**arkov **M**odel
i	**i**nversé
IR	**I**nverted **R**epeat (Répétition Inversée)
IS	**I**nsertion **S**equence
IUPAC	International Union of Pure and Applied Chemistry
kb	**kilob**ases
LARD	**LA**rge **R**etrotransposon **D**erivative
LINE	**L**ong **IN**terspersed repetitive **E**lement
LTR	**L**ong **T**erminal **R**epeat
MAPK	**M**itogen-**A**ctivated **P**rotein **K**inase
Mb	**M**éga**b**ase
MITE	**M**iniature **I**nverted **T**ransposable **E**lement

MULE	**MU**tator-**L**ike **E**lements
NASC	**N**ottingham **A**rabidopsis **S**tock **C**entre
NP	**N**on **P**olynomial
ORF	**O**pen **R**eading **F**rame
pol	**pol**ymérase
pollen-RE	**pollen r**esponsive **e**lement
pb	**p**aire de **b**ases
RAG	**R**ecombination **A**ctivating **G**ene
RNAi	**RNA i**nterférence
RPA	**R**eplication **P**rotein **A**
RTE	**R**etro**t**ransposable **E**lement
SINE	**S**hort **In**terspersed repetitive **E**lement
STAN	**S**uffix **T**ree **AN**alyser
SVG	**S**calable **V**ector **G**raphic
SVG	**S**tring **V**ariable **G**rammar
TAIR	**T**he **A**rabidopsis **I**nformation **R**essource
TIR	**T**erminaison **I**nversée **R**épétée
tnpA	transposase
tnpR	résolvase
TRIM	**T**erminal-**R**epeat retrotransposon **I**n **M**iniature
TSD	**T**arget **S**ite **D**uplication
UTR	**U**nTranslated **R**egion
VDJ	**V**ariable **D**iversity **J**oining

Introduction

Contexte de la thèse

Un génome est l'ensemble des séquences d'ADN (Acide DésoxyriboNucléique) présentes dans une cellule ou dans ses structures subcellulaires. Une plante contient ainsi un génome nucléaire, un génome mitochondrial et un génome plastidial. Notre objet d'étude est le génome nucléaire des eucaryotes et plus particulièrement celui de la plante modèle *Arabidopsis thaliana*.

Depuis le séquençage du bactériophage $\Phi X174$ en 1977 [180] et le séquençage du premier organisme vivant : la bactérie *Haemophilus influenzae* [72], le séquençage des génomes a permis de mieux appréhender le fonctionnement du vivant grâce à l'annotation et à l'étude des gènes des différents organismes.

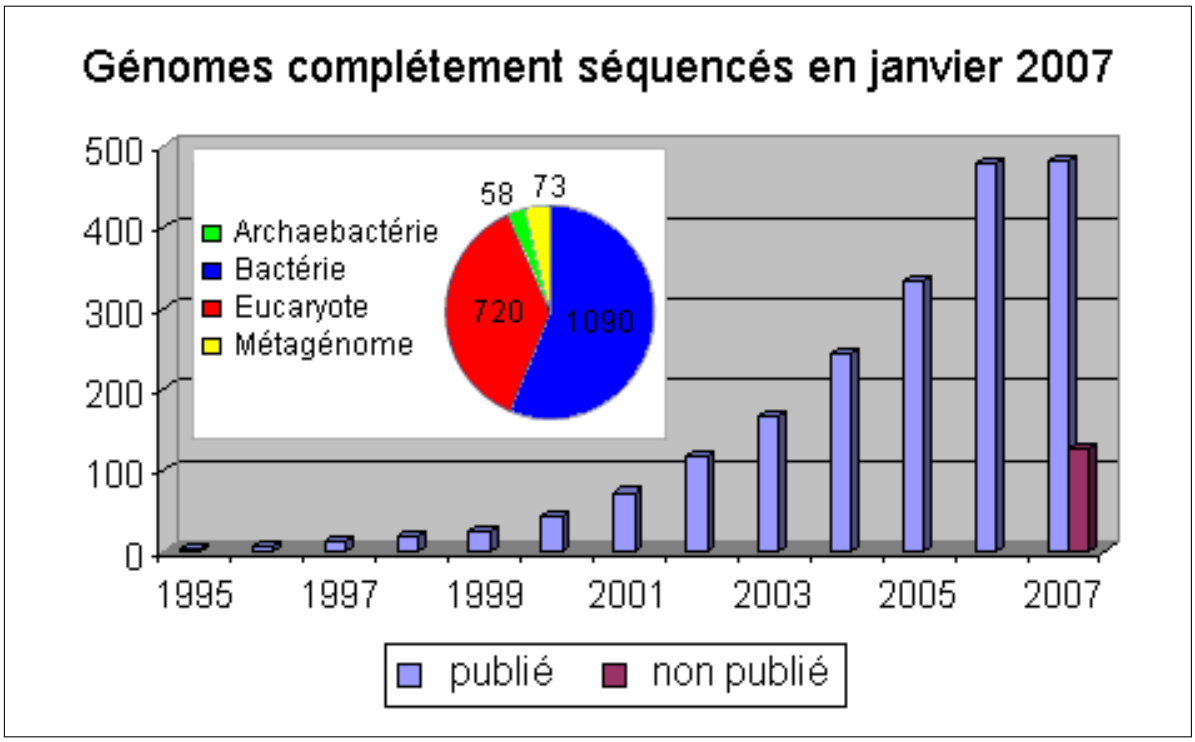


FIG. 1.1 – Nombre de génomes séquencés au cours des années 1995-2006 [118, 132]. Le camembert en haut à gauche indique le nombre de génomes séquencés selon les quatres règnes du vivant.

Le site Web Genomes OnLine Database (www.genomesonline.org) référence tous les

génomes étudiés (Figure 1.1). Il dénombre actuellement 2469 génomes séquencés ou en cours de séquençage [118, 132].

La stratégie usuelle de séquençage d'un génome est la méthode du Shotgun. Cette méthode consiste à fragmenter le génome, et à insérer les fragments dans des vecteurs (généralement des BAC (Bacterial Artificial Chromosome)). On obtient ainsi la séquence nucléotidique découpés aléatoirement dans ses parties en petites portions (Figure 1.2) [83]. La séquence d'origine est ensuite reconstituée grâce à un programme d'alignement qui permet l'assemblage des séquences.

A l'exception d'une souche de bactérie, tous les organismes séquencés ou en cours de séquençage présentent dans leur génome des éléments répétés appelés "éléments transposables". Ces éléments, répétés en proportion variable, posent des problèmes d'assemblage des séquences. En effet, l'assemblage est effectué sur la base de la reconnaissance de séquences chevauchantes ayant une large portion similaire. Or les éléments répétés se confondent facilement avec les répétitions dues à la réplication en de nombreux exemplaires du matériel génétique que demande la stratégie de séquençage en Shotgun.

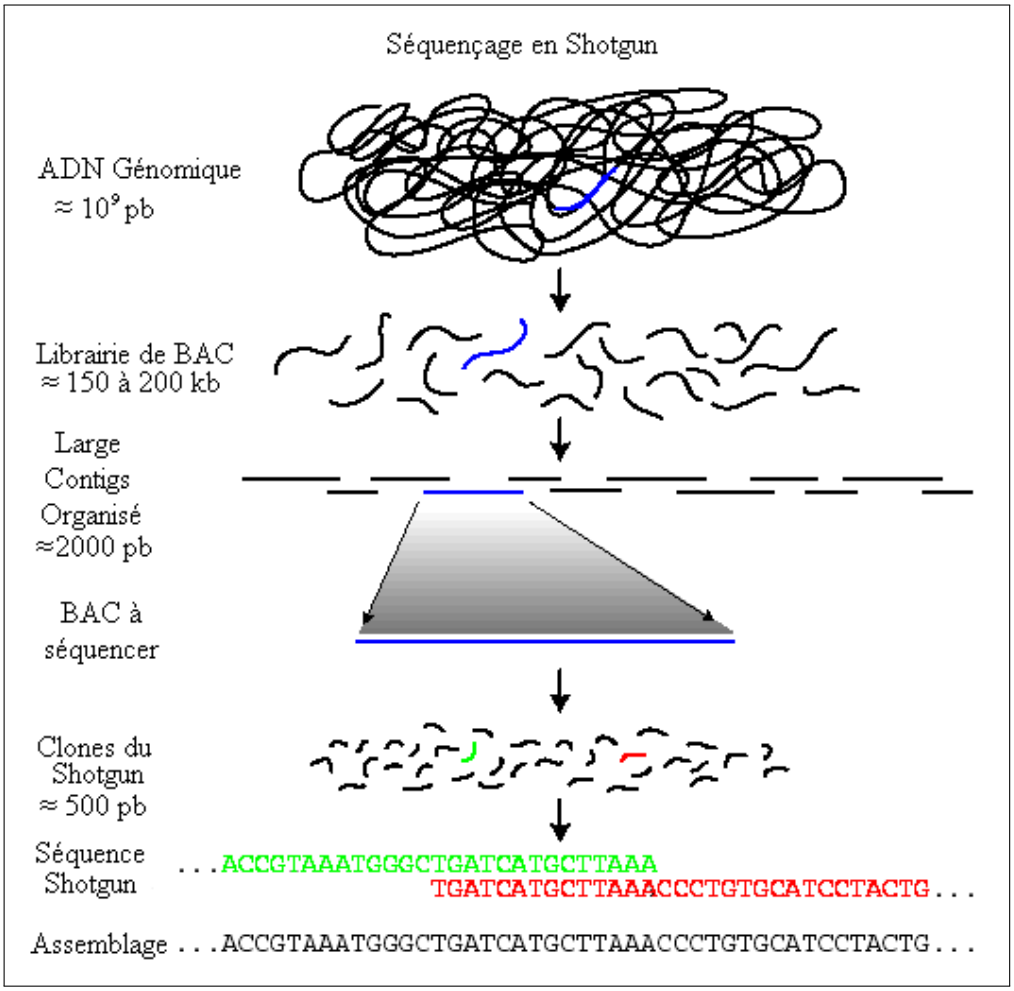


FIG. 1.2 – Représentation schématique de la méthode de séquençage et d'assemblage d'un génome [83].

De nombreux mécanismes contribuent à l'évolution des génomes. Les éléments transposables font partie des principaux acteurs de cette variabilité appelée "dynamique des

génomes". Ce projet de thèse repose sur l'étude de ces éléments transposables et l'analyse de la dynamique évolutive qu'ils exercent sur le génome.

L'introduction de cette thèse est composée principalement de trois sections afin de préciser nos sujets d'étude : les mécanismes de variabilité génomique, les éléments transposables et une classe particulière d'éléments transposables sur laquelle nous avons principalement travaillé, les hélitrons. Ces parties nous permettront d'introduire le vocabulaire biologique nécessaire pour la compréhension de cette thèse mais surtout d'appréhender l'importance des éléments transposables dans la dynamique des génomes. Le chapitre suivant présentera les différentes méthodes bioinformatiques utilisées et permettra de cerner les possibilités et les difficultés rencontrées lors de l'étude *in silico* (c'est à dire par traitement informatique des séquences) des éléments transposables dans un génome entier. Le chapitre "Résultats et Discussion" développera l'ensemble des résultats biologiques et bioinformatiques obtenus, et notre contribution pour une meilleur compréhension de la dynamique des génomes. Enfin, le dernier chapitre abordera différentes approches possibles qui peuvent être explorées à la suite de cette thèse.

1.1 Mécanismes à l'origine d'une variabilité du génome

L'analyse de la séquence des génomes a montré une grande variabilité de séquences inter-espèces. L'étude de la génétique des populations a aussi montré une grande variabilité entre individus d'une même espèce. Ceci est en partie dû au mécanisme de variabilité le plus connu : la mutation ponctuelle d'un nucléotide (incluant l'insertion, la délétion ou la substitution). Néanmoins, les mécanismes de mutation ponctuelle ne seront pas abordés au cours de ce rapport. En effet, les éléments transposables, sujet principal de cette thèse, ne provoquent pas ce type de mutation, mais des mutations de séquences plus importantes.

En plus des éléments transposables, il existe d'autres mécanismes de mutation qui modifient des portions entières de la séquence d'un génome et représentent des facteurs importants d'évolution et d'adaptation d'un organisme ou d'une espèce. Dans un premier temps, nous allons aborder les mécanismes de variabilité génomique qui agissent sur certains types de cellules ou à certains moments de la vie de la cellule puis nous aborderons les autres mécanismes de variabilité d'un génome. Tous ces mécanismes nous permettront d'appréhender la dynamique d'un génome et de positionner l'intervention des éléments transposables.

1.1.1 Mécanismes de variabilité spécifiques

Beaucoup de phénomènes de variation n'affectent pas l'ensemble des cellules de l'organisme ou ne se transmettent pas à la génération suivante. Nous nous concentrerons ici sur les mécanismes qui n'affectent que certaines cellules de l'organisme, ou ne sont actifs qu'à un moment précis de la vie de la cellule. L'exemple le plus frappant est la recombinaison VDJ (Variable Diversity Joining : nom des trois gènes qui participent à cette recombination) des lymphocytes (voir ci-dessous). Ce mécanisme permet à l'organisme de modifier le génome de quelques-unes de ces cellules pour répondre spécifiquement aux attaques des pathogènes [87].

1.1.1.1 Variabilité liée à certains types cellulaires : la recombinaison VDJ

La recombinaison VDJ se déroule dans les cellules immunitaires des mammifères [87]. Lors d'une attaque par un pathogène, l'organisme crée des récepteurs et des anticorps (immunoglobulines) spécifiques à ce pathogène. Ainsi l'organisme, qui ne possèdent par le gène codant dans son génome, produit une protéine unique qui se lie spécifiquement à l'organisme étranger[87]. Dans les futurs lymphocytes T, les gènes V (Variable) et C (Constant) sont distants de plusieurs milliers de nucléotides. Le complexe RAG1 et RAG2 (Recombination Activating Gene) qui est formé de protéines dérivées d'éléments transposables [100] reconnaît un motif palindromique au niveau des gènes D (Diversity) et J (Joining). Il excise l'ADN entre les deux gènes (mécanisme similaire à la transposition), puis DJ est rapproché du gène V grâce au même mécanisme (Figure 1.3). La transcription et l'épissage éliminent les introns entre les différents domaines pour créer la chaîne lourde du récepteur protéique des lymphocytes T. L'appellation chaîne lourde et chaîne légère fait référence au poids moléculaire plus faible de la chaîne légère par rapport à la chaîne lourde. La même recombinaison a lieu pour la chaîne légère mais elle ne contient pas de domaines D (Figure 1.3). Le même processus se déroule dans les lymphocytes B avec les

chaînes légères et lourdes des immunoglobulines [87]. La cellule, contenant la nouvelle protéine, est mise en contact avec le pathogène. Si la protéine a une grande affinité avec le pathogène, la cellule se multipliera et sera distribuée dans l'organisme, sinon elle sera détruite.

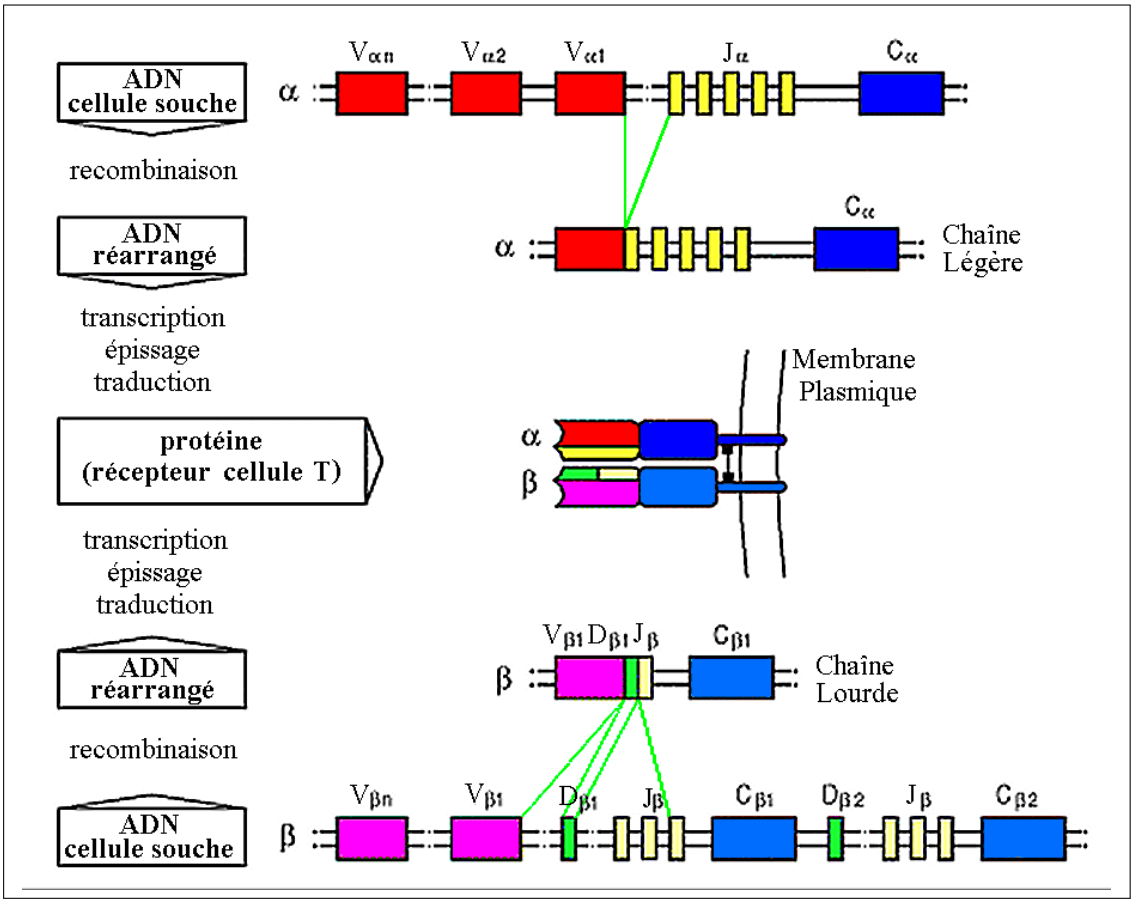


FIG. 1.3 – Recombinaison VDJ pour la formation de la protéine réceptrice du lymphocyte T [87]. Les gènes TCR α (chaîne légère) et β (chaîne lourde) sont composés de domaines discontinus qui sont réarrangés durant le développement de la cellule T. La chaîne légère est composée des fragments V et J qui sont réarrangés. Comme les chaînes lourdes d'immunoglobuline, la chaîne β est composée des fragments V, D et J. Ces segments sont réarrangés dans la chaîne lourde β. Les réarrangements, l'épissage puis la traduction de l'ARNm (Acide RiboNucléique messager) créent la chaîne α et β de la protéine réceptrice des lymphocytes T.

La base de données IMGT ((imgt.cines.fr/) recense les séquences d'immunoglobuline de 150 espèces dont 1512 gènes uniquement dédiés aux gènes d'immunoglobuline et aux récepteurs des lymphocytes T présents chez la souris et l'être humain [125]. Cette base est associée à de nombreux outils classiques d'identification ainsi qu'un outil de prédiction des recombinaisons VDJ (IMGT/JunctionAnalysis).

Néanmoins, cette dynamique du génome ne concerne pas toutes les cellules de l'organisme mais seulement quelques cellules et dans des situations bien particulières. D'autres mécanismes interviennent sur toutes les cellules de l'organisme, mais uniquement à des stades particuliers du développement de la cellule tels que les divisions cellulaires avec les

recombinaisons inégales.

1.1.1.2 Variabilité liée à certains stades de développement : les recombinaisons inégales

A la différence de la recombinaison VDJ qui diminue le nombre de nucléotides du génome, les recombinaisons inégales peuvent augmenter ou diminuer le nombre de nucléotides. Quand les deux chromosomes d'un génome diploïde se séparent au cours d'une méiose (division sexuée de la cellule) ou d'une mitose (duplication non sexuée de la cellule), les recombinaisons inégales permettent aux deux cellules filles d'obtenir un génome différent de la cellule mère [21]. La recombinaison inégale est principalement produite par trois phénomènes biologiques : les crossing-over inégaux, les échanges inégaux entre chromatides sœurs et les recombinaisons inégales pendant la réplication de l'ADN (Figure 1.4).

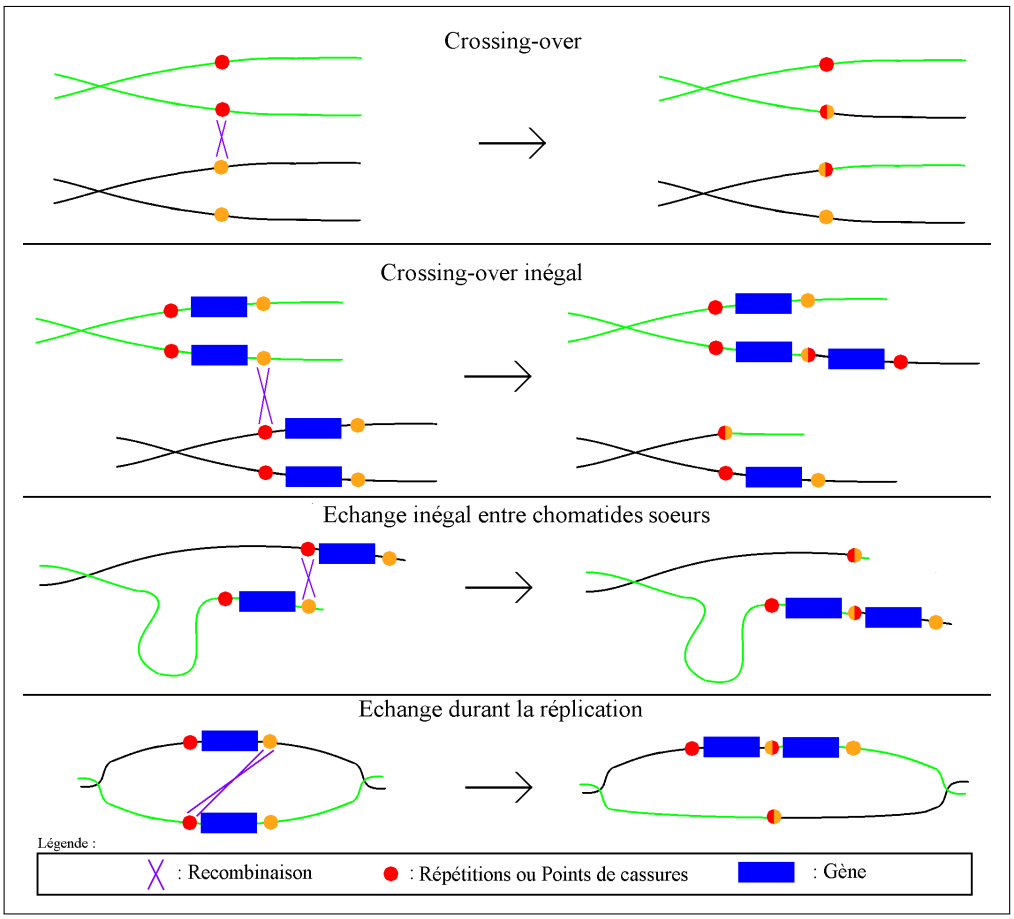


FIG. 1.4 – Recombinaisons inégales de l'ADN [21].

Ces trois recombinaisons inégales sont toutes dues à des éléments répétés (par exemple des éléments transposables) présents dans le génome [21]. Au cours de la méiose ou de la mitose, les deux chromosomes homologues peuvent échanger des portions de séquences grâce à deux cassures chromosomiques qui permettent de maintenir le même nombre nucléotidique dans chaque brin. Dans une recombinaison inégale, il y a une cassure au niveau d'une répétition, à une position donnée sur un chromosome, et une autre cassure

dans la même répétition, mais à une position différente sur un autre chromosome (Figure 1.4) [21]. La recombinaison entraîne l'échange des deux brins chromosomiques après la cassure.

Pour les crossing-over entre deux chromosomes homologues ou pour les échanges inégaux entre chromatides sœurs (deux brins du même chromosome), une différence de taille est observée entre les chromosomes qui ont subi cette recombinaison et les chromosomes intacts. Cette différence est due à la différence de position de l'élément répété utilisé lors de la recombinaison. Le brin chromosomique dont l'élément répété est le plus éloigné du centromère (portion du chromosome utilisée lors de la séparation des chromosomes dans la méiose) augmente de taille, alors que le brin dont l'élément répété est le plus proche du centromère diminue de taille (Figure 1.4). Par exemple, sur le chromosome 4 du génome d'*Arabidopsis thaliana*, le taux moyen de crossing-overs est de 4,6 centiMorgans par Mégabase (Mb) [54]. Un centiMorgan correspond à 1 crossing-over pour 100 méioses. Les éléments transposables sont responsables de 15 à 30 % de ces crossing-overs chez *Arabidopsis thaliana* [153].

Concernant l'échange d'ADN durant la réplication, l'une des deux copies du chromosome perd la partie de séquence présente entre les deux copies de l'élément répété responsable de cette cassure (Figure 1.4). L'autre copie acquiert la partie entre ces deux éléments répétés. Si la partie génomique entre les deux répétitions contient un gène, l'une des cellules filles contiendra les deux copies du gène alors que l'autre cellule fille aura perdu ce gène (Figure 1.4). Cette recombinaison est surtout présente dans les génomes bactériens et au cours de la méiose des génomes eucaryotes [21]. Chez l'être humain, le taux de recombinaisons inégales estimé varie par génération de 10^{-10} à 10^{-9} [108]. Ces recombinaisons peuvent provoquer des maladies génétiques telles que des dystrophies musculaires [151].

Les recombinaisons inégales sont détectées grâce à l'alignement des deux séquences cibles ayant subi la recombinaison avec une séquence de référence. La comparaison avec la séquence de référence montre qu'une délétion apparaît dans la séquence recombinée qui a perdu un locus, et qu'une insertion est observée dans la séquence recombinée qui a gagné un locus. La séquence insérée doit être identique à la séquence délétée pour prouver qu'il s'agit bien d'une recombinaison inégale.

1.1.2 Mécanismes de variabilité générale du génome

D'autres mécanismes sont capables de provoquer des modifications de séquence plus générales dans tous les types cellulaires et à tous les instants de leur cycle biologique. Ces mécanismes sont principalement réalisés par les virus intégrés aux génomes, les éléments transposables, les mini/microsatellites et les transferts horizontaux. Ces éléments sont donc des facteurs importants dans la dynamique de variation d'un génome. Les mini/microsatellites se différencient des éléments transposables et des virus par leur manque de mobilité : les satellites sont des répétitions locales alors que les insertions virales et les éléments transposables sont dispersées dans le génome. Nous établissons aussi dans cette partie que les éléments transposables ont une action indirecte sur les autres mécanismes globaux de variation du génome, soulignant encore leur importance et la nécessité de les étudier pour comprendre la dynamique des génomes.

1.1.2.1 Les microsatellites et les minisatellites

Les satellites sont de courtes séquences répétées en tandem (l'une à la suite de l'autre). Les microsatellites sont des séquences répétées de 1 à 5 nucléotides alors que le motif des minisatellites peut atteindre une taille de 100 nucléotides [21]. Certaines régions du génome humain contiennent plusieurs centaines, voire plusieurs milliers d'unités répétées en tandem [21, 194]. Si les micro/minisatellites sont répartis dans tout le génome, certaines régions telles que les télomères présentent une plus grande concentration de ces éléments [66]. En effet, un micro/minisatellite est créé au cours de la réplication de l'ADN. Parfois l'ADN polymérase est bloquée par une structure secondaire de l'ADN et "patine" sur cette portion d'ADN : la polymérase revient en arrière de un à plusieurs nucléotides, relit la séquence et la recopie en créant une nouvelle unité du satellite sur le brin nouvellement créé [21]. Ce retour peut être répété plusieurs fois au cours d'une même réplication. Cette structure secondaire de l'ADN provient des séquences déjà répétées en tandem ou de séquences palindromiques (par exemple les palindromes des éléments transposables). Les satellites ont peu de rôles fonctionnels connus [21]. Comme les recombinaisons inégales, les micro/minisatellites sont responsables de maladies comme le syndrome de Lynch chez l'être humain [88].

1.1.2.2 Les virus

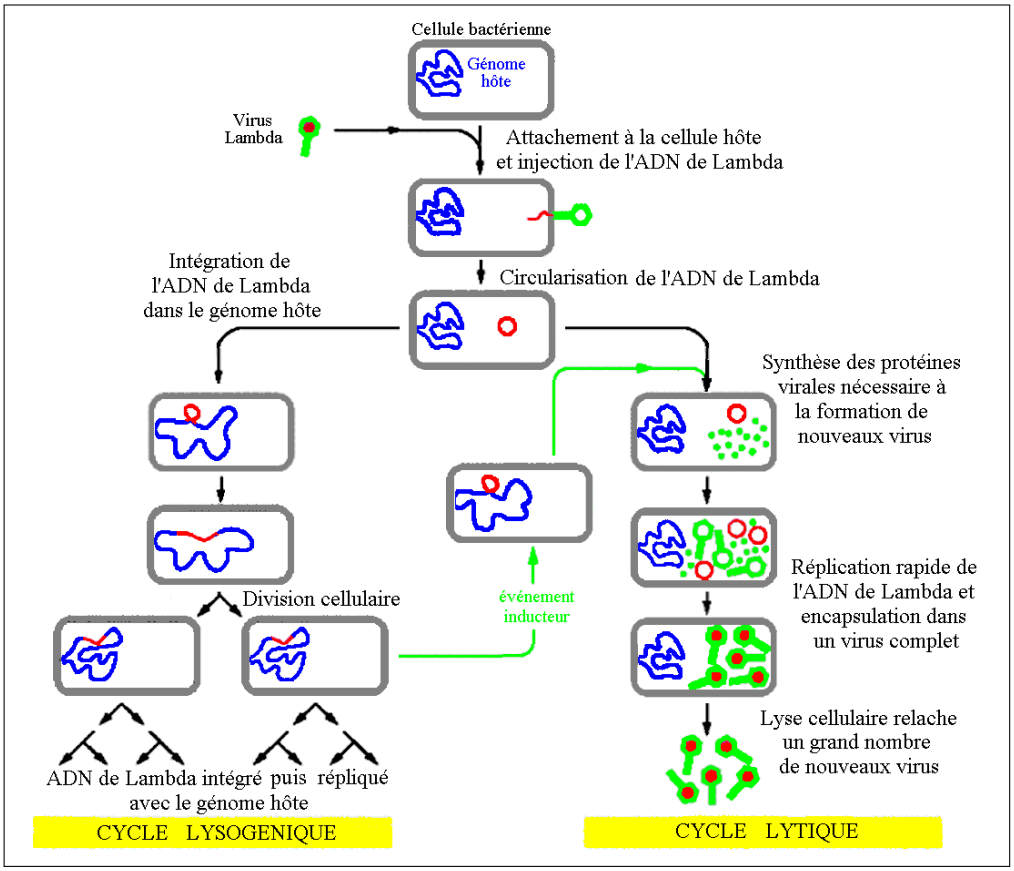


FIG. 1.5 – Différents cycles de réplication d'un virus Lambda dans une bactérie [22].

Les virus sont des éléments génétiques mobiles. Ils sont considérés comme des parasites

moléculaires et non comme des organismes vivants, car ils sont incapables de se répliquer seuls [22]. La Figure 1.5 décrit le cas d'un virus à ADN. Le virus se fixe à la cellule hôte et injecte son génome (Figure 1.5). Après circularisation de l'ADN, le virus peut agir selon deux modes de réplication : le cycle lytique et le cycle lysogénique (Figure 1.5) [22]. Le cycle lytique permet au virus de se multiplier rapidement dans la cellule hôte jusqu'à sa mort (la lyse) (Figure 1.5). Le cycle lysogénique quant à lui préserve la cellule hôte. En effet, l'ADN viral s'intègre dans le génome hôte et devient silencieux pour l'hôte. Au cours de sa réplication, la cellule copie aussi le génome viral intégré.

Les virus peuvent occuper une large proportion du génome hôte, par exemple le génome humain est composé de 1 à 2 % de rétrovirus *HERV* [144]. Un virus peut encapsider (par erreur) une partie du génome hôte, cette séquence est alors transportée dans un autre organisme et permet le transfert horizontal de gènes par voie virale. Ainsi, les virus parasitent continuellement tous les types d'organismes vivants, ce qui leur confèrent un rôle important dans l'évolution des espèces et la dynamique des génomes.

De nombreux virus sont apparentés à des éléments transposables : par exemple les rétrovirus et les rétrotransposons ont des séquences nucléotidiques similaires à l'exception du gène de l'enveloppe virale absent chez les éléments transposables [39].

1.1.2.3 Les transferts horizontaux

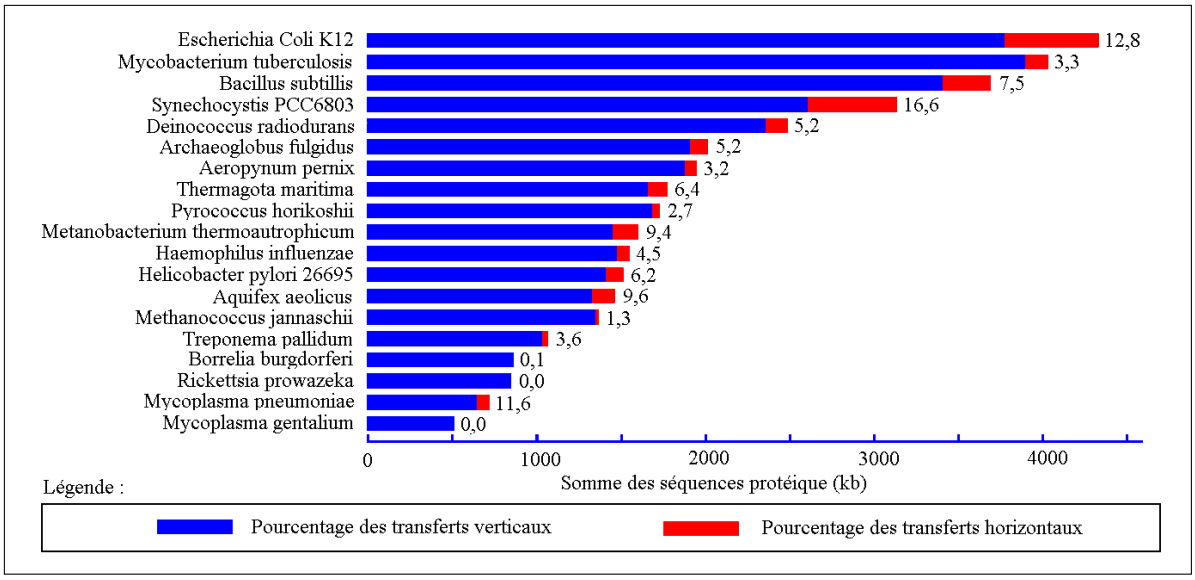


FIG. 1.6 – Proportion de transferts génétiques horizontaux dans certains génomes bactériens [21].

Les transferts horizontaux sont des transferts d'ADN d'un organisme à un organisme d'une autre espèce [21]. Ces transferts permettent l'acquisition de nouveaux gènes par le génome receveur. Ces transferts sont opérés par les éléments génétiques mobiles tels que les virus et les éléments transposables.

Les transferts horizontaux ont une place très importante dans l'évolution et la variabilité des procaryotes. Par exemple plus de 10 % des séquences codantes d'*Escherichia coli* proviennent de transferts horizontaux (Figure 1.6) [21]. Les transferts horizontaux

sont plus difficiles à montrer chez les organismes eucaryotes, mais la phylogénie (étude de l'évolution des gènes) de certains gènes permet d'en identifier les traces [4]. Il ne s'agit bien sûr que d'observations indirectes, par analyse comparative des génomes contemporains. La phylogénie permet de produire des hypothèses de transferts au cours de l'évolution expliquant la variabilité observée.

L'analyse des précédents mécanismes de variabilité ponctuelle ou générale des génomes a mis en évidence l'implication directe ou indirecte des éléments transposables (ETs). Ces éléments transposables sont aussi eux-mêmes des acteurs importants de variabilité, de par leur activité de transposition et de par leur capacité d'action dans tous les types cellulaires et à tous les moments du développement cellulaires. Nous allons maintenant établir le rôle direct et majeur des ETs dans la plasticité des génomes, en décrivant leurs caractéristiques structurales et fonctionnelles et leur mode d'interaction sur les génomes.

1.2 Les éléments transposables

Contexte historique

Barbara McClintock a mis en évidence en 1944 des phénomènes d'instabilité génétique chez le maïs qu'elle attribue à l'élément mobile Ds (Dissociation) et à l'élément Ac (Activateur). Ce n'est qu'en 1951 qu'elle présente pour la première fois les éléments transposables au Cold Spring Harbor Symposium [142]. Elle obtiendra le prix Nobel en 1983 pour ses travaux sur les éléments transposables (ETs).

Chez les bactéries, les séquences IS (Insertion Sequence) sont découvertes en 1969 dans l'*opéron gal* de *Escherichia coli*. Au cours des 30 dernières années, de nouvelles familles d'éléments transposables ont été découvertes dans tous les génomes eucaryotes et procaryotes connus. Aujourd'hui encore, de nouveaux types d'éléments transposables sont découverts. Par exemple, le type Polintons/Maverick [96] est le dernier type d'éléments transposables décrit en mars 2006.

1.2.1 Définition des éléments transposables

Les éléments transposables (ETs) sont des séquences d'ADN répétées, caractérisées par leur capacité à se multiplier au sein du génome hôte par des processus moléculaires plus ou moins identifiés, rassemblés sous le terme de transposition [39]. La transposition est définie par le mouvement de matériel génétique d'une position chromosomique à une autre. Il est important de restreindre cette notion à la capacité que possèdent certains segments d'ADN à se déplacer vers de nouveaux sites génomiques par des mécanismes indépendants du processus de recombinaison homologue. Les séquences qui possèdent cette capacité intrinsèque à changer de localisation chromosomique sont appelées éléments mobiles ou éléments transposables [39].

1.2.1.1 Structure générale des éléments transposables

Les éléments transposables sont bordés par des répétitions directes ou TSD (Target Site Duplication) qui sont généralement créées lors de l'insertion de l'élément transposable. Les TSDs sont caractéristiques d'une famille d'éléments transposables. La plupart des ETs ont aussi à leurs extrémités des répétitions inversées terminales (Figure 1.7). La taille de ces répétitions inversées dépend de la famille de l'élément transposable.

On distingue deux types d'éléments transposables : les éléments transposables autonomes, et les ETs non-autonomes. Les éléments transposables autonomes contiennent un ou plusieurs ORFs codant pour des protéines fonctionnelles impliquées dans sa transposition et la régulation de cette transposition (Figure 1.7). Les ETs non-autonomes ne possèdent pas d'ORFs (Open Reading Frame) correspondant à ces protéines de transposition. Les éléments transposables sont répartis en plusieurs catégories selon leur mode de transposition [71].

1.2.1.2 Mécanismes de transposition et classification des éléments transposables

Les éléments de classe I, ou rétrotransposons, transposent par transcription de l'ADN en ARN simple brin, lui-même réverse-transcrit en ADN et inséré dans le génome hôte

(Figure 1.7). Les rétrotransposons présentent un mécanisme de transposition de type "copier-et-coller" [71]. Les éléments de classe II, ou transposons, utilisent un intermédiaire de type ADN et assurent leurs transpositions selon un processus de type "couper-et-coller" (Figure 1.7).

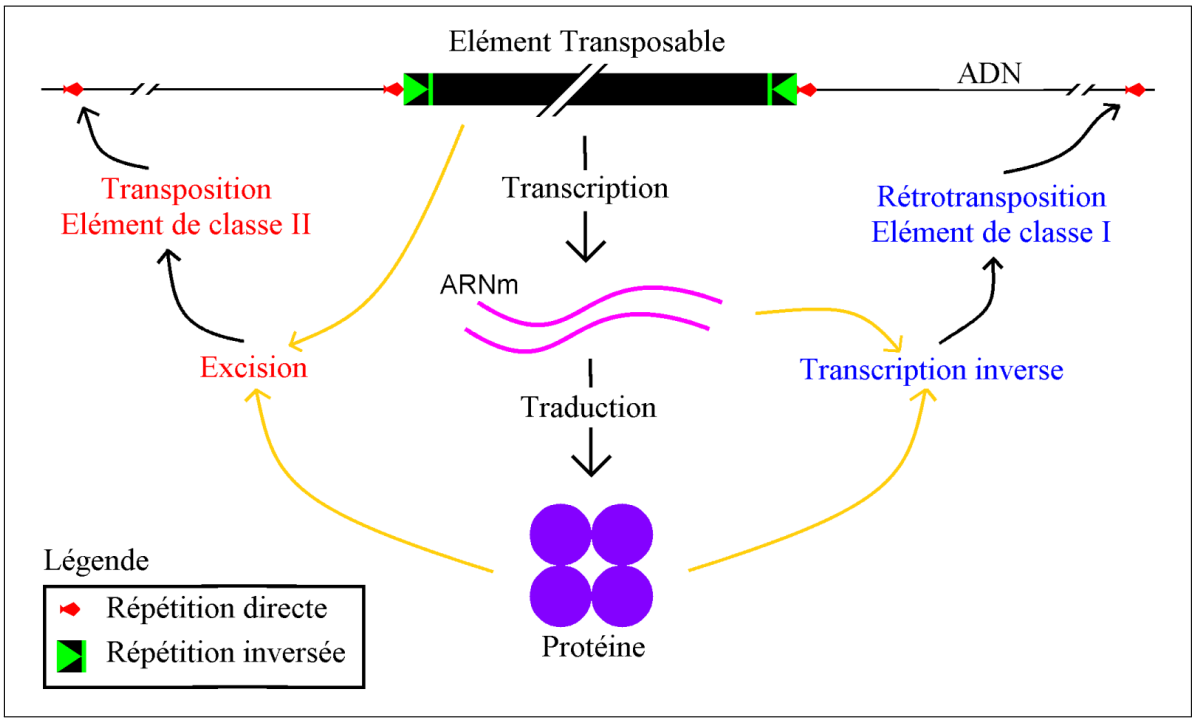


FIG. 1.7 – Mécanismes généraux de transposition. Les ORFs des ETs autonomes sont transcrits puis traduits. Les protéines codées par les éléments de classe I vont reconnaître la molécule d'ARN pour réaliser la rétrotransposition. Les protéines codées par les éléments de classe II reconnaissent la molécule d'ADN et la coupent pour réaliser la transposition.

Dans chacune de ces deux classes d'éléments transposables, on peut définir des sous-classes couramment décrites en termes de superfamille et de famille d'ETs. Une famille d'éléments transposables est un ensemble de séquences répétées, présent dans le même organisme, et partageant de fortes caractéristiques telles que les ORFs ou la séquence entière. Une superfamille d'éléments transposables est un ensemble de familles d'éléments transposables partageant des caractéristiques communes (généralement moins strictes que celles d'une famille), et présent dans un ou plusieurs génomes. Par exemple, la superfamille Tc1/mariner est l'ensemble des familles Tc1/mariner réparties dans les génomes eucaryotes qui partagent les mêmes répétitions inversées et les mêmes TSD.

1.2.2 Les éléments transposables de classe I ou rétrotransposons

Les éléments de classe I sont apparentés aux virus [39] et, comme eux, contribuent largement à la dynamique des génomes. Ces éléments sont aussi appelés rétrotransposons à cause de leur similarité structurale avec le génome des rétrovirus. De plus, ils représentent la majorité des éléments transposables chez les mammifères et parfois une large proportion de leur génome. Par exemple, ils représentent plus de 45 % du génome humain [83].

Grâce à leur mécanisme de transposition (”copier-et-coller”), une seule matrice d'ADN peut donner plusieurs centaines de copies au cours d'un seul événement de transposition. Ces éléments possèdent une capacité d'invasion très importante au sein des génomes hôtes. De ce fait, ils constituent une grande partie des génomes eucaryotes [114]. Les rétrotransposons semblent n'exister que dans ces génomes eucaryotes. Les rétrotransposons sont eux-mêmes classés en deux sous-classes : les rétrotransposons à LTR (Long Terminal Repeat) et sans LTR.

1.2.2.1 Rétrotransposons à LTR

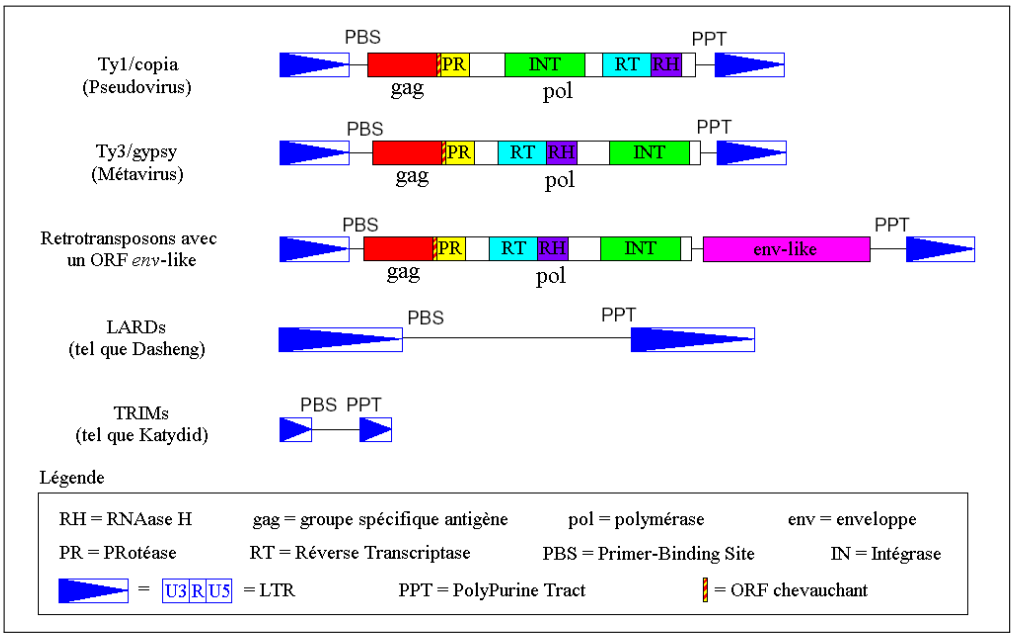


FIG. 1.8 – Organisation génomique des différents rétrotransposons à LTR [79, 176].

Les LTRs (5' et 3'), longs de 100 pb à 5 kb, possèdent des répétitions inversées à chaque extrémité qui se divisent en trois parties : U3, R et U5 [114, 176] (Figure 1.8). La partie U3 contient l'enhancer (motif de régulation activant un promoteur) et le promoteur nécessaires et suffisants pour la transcription du rétrotransposon. La partie R contient l'hairpin (tige-boucle) qui sert de signal de début de la réverse-transcription. La partie U5 contient le site de polyadénylation nécessaire pour finir la transcription. Les rétrotransposons autonomes possèdent deux ORFs codant pour les gènes *gag* et *pol*. Le gène *gag* (Group Associated Antigen) code une polyprotéine à l'origine de la matrice, de la capside et de la nucléocapside des particules de type rétroviral. Le gène *pol* (POLymérase) contient des domaines protéiques pour une protéase, une transcriptase inverse qui convertit la molécule d'ARN en ADN double brin, une RNAaseH qui dégrade l'ARN lors de la synthèse du deuxième brin d'ADN, et une intégrase responsable de l'intégration de la nouvelle copie dans le génome hôte (Figure 1.8). Après la traduction de l'ARNm (ARN messager) en séquence primaire de protéines, cette séquence subit un clivage protéolytique qui donnera naissance aux différentes protéines actives.

Lors d'un événement de rétrotransposition, la séquence d'ADN est transcrite à partir de la partie R présente dans le LTR 5' jusqu'à la partie R du LTR 3' [114]. L'ARNm

résultant est traduit dans le cytoplasme en protéines nécessaires pour la réverse transcription et l'intégration. Un ARNt (ARN de transfert) s'associe avec l'ARNm et forme alors une courte région ARN double brin, qui sert de point d'ancrage à la reverse transcriptase. Après la formation d'un complexe ADN-ARN, la RNAaseH détruit cet ARN pour laisser l'ADN simple brin. Cet ADN s'hybride avec une structure circulaire qui permet de continuer la réverse transcription jusqu'à l'obtention d'un ADN simple brin complet et circulaire. Le deuxième brin d'ADN est initié au niveau de la zone polypurine (PolyPurine Tract (PPT)) proche du 3' LTR. L'ADNc (ADN complémentaire) double brin obtenu est ensuite intégré dans le génome hôte [79, 27].

Les rétrotransposons sont sous-classés en deux groupes distincts en fonction de la position des ORFs présents dans le gène *pol* : les Ty1-copia et les Ty3-gypsy [176] (Figure 1.8). La principale différence entre ces deux groupes vient de l'ordre des protéines présentes dans le gène *pol*. Chez Ty1-copia l'intégrase est située avant la réverse transcriptase et la RNAaseH ; chez Ty3-gypsy, la réverse transcriptase et la RNAaseH sont situés avant l'intégrase. Les Ty3-gypsy sont apparentés aux *Metavirus*, les Ty1-copya sont apparentés soit aux *Pseudovirus* comme Ty1 soit aux *Hemivirus* comme Ty5 (ces deux dernières familles appartiennent à la superfamille *Pseudovirae*) [39].

Récemment, des rétrotransposons à LTR non-autonomes, tels que les LARDs (LArge Retrotransposon Derivative) ou les TRIMs (Terminal-Repeat retrotransposon In Miniature), ont été découverts dans les génomes de plantes (Figure 1.8) [176] . Les TRIMs sont les plus courts rétrotransposons à LTR : 500 bp avec des LTRs mesurant de 100 à 250 bp. Leur mécanisme de transposition semble lié aux éléments autonomes, mais aucune transposition n'a été décrite à ce jour. Les LARDs possèdent de larges LTRs, jusqu'à 4,5 kb et une région interne riche en structures secondaires stables.

1.2.2.2 Rétrotransposons sans LTR ou rétroposons

Les rétrotransposons sans LTR, aussi appelés rétroposons, ne possèdent pas de répétitions terminales inversées [71, 176]. L'extrémité 3' présente une queue polyA de taille variable et un signal de polyadénylation. Contrairement aux rétrotransposons à LTR, les rétroposons n'utilisent pas d'ADN cytoplasmique au cours de leur transposition [176]. La grande diversité des rétroposons et leur similitude avec les rétrotransposons à LTR suggèrent que les rétroposons seraient à l'origine des rétrotransposons à LTR [39].

Les rétroposons transposent via un mécanisme commun aux introns de groupe II appelé TRRT (Target-primed Reverse Transcription) [156, 39]. L'endonucléase réalise une coupure simple brin dans l'ADN de l'hôte. La queue polyA de l'ARN du rétroposon se fixe à un ADN polyT simple brin. L'ARN du rétroposon sert d'amorce pour la synthèse de l'ADNc du rétroposon par la transcriptase inverse. Une deuxième coupure intervient alors sur l'autre brin d'ADN de l'hôte permettant l'intégration finale du rétroposon sous forme ADN [156, 158]. Les rétroposons autonomes contiennent deux ORFs : le premier codant une protéine similaire à la protéine gag des rétrotransposons à LTR, le second ORF codant au moins une endonucléase et une réverse transcriptase (Figure 1.9). La plupart de ces ETs peuvent être classés en quatre familles appelées LINE (Long INterspersed repetitive Element), SINE (Short Interspersed repetitive Element), élément I ou RTE (Retrotransposable Element) [134].

Les RTEs, récemment découverts, sont des rétroposons autonomes d'environ 3,3 kb et possèdent de larges répétitions d'une centaine de paires de bases [134]. Contrairement

aux autres rétroposons, l'élément RTE ne code que pour l'ORF2 (Figure 1.9), ce qui en fait le plus petit rétroposon autonome connu actuellement.

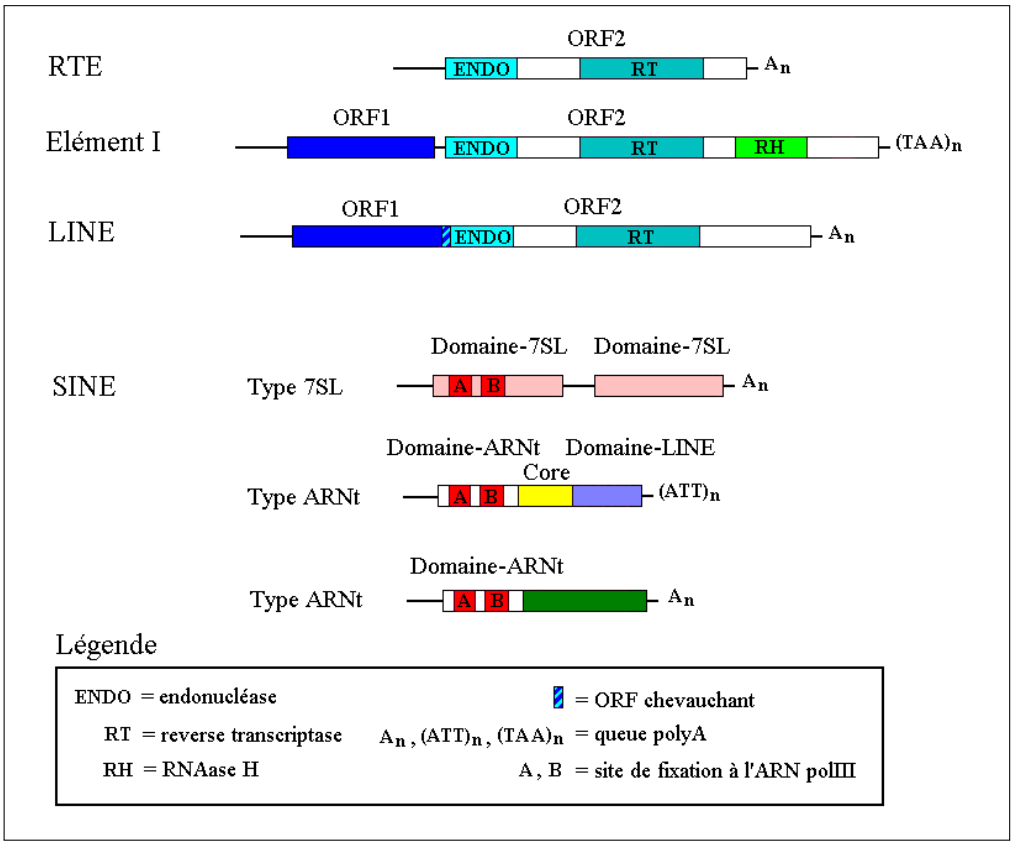


FIG. 1.9 – Organisation structurelle des rétroposons [49].

Les éléments I mesurent environ 5,4 kb. En plus de l'endonucléase et de la reverse transcriptase, l'ORF2 code pour une RNAaseH (Figure 1.9). Il possède une queue polyA en 3' particulière composée d'une succession de nucléotides TAA et ne contient pas de signal de polyadénylation en 3' [39]. Cet élément transpose spécifiquement dans les ovocytes des drosophiles, provoquant ainsi le phénomène de dysgénèse des hybrides (si on croise un mâle avec l'élément I et une femelle sans cet élément, on obtient une très forte mortalité embryonnaire) [39].

La séquence interne d'un élément LINE (Ex : famille L1) est composée de deux ORFs, d'une queue polyA de longueur variable, d'un 5' UTR (UnTranslated Region = région non traduite), et d'un 3' UTR d'environ 670 pb [39]. L'ORF1 code une protéine de liaison à l'ARN (Figure 1.9). La plupart des LINEs sont non fonctionnels car les 2 ORFs sont saturés de codons STOP ou de décalages du cadre de lecture. La taille des LINEs varie de 5 kb à 7 kb pour les éléments fonctionnels [79, 156] et peut atteindre 1 kb pour les éléments non-autonomes.

Les SINEs (tels que la famille Alu dans le génome humain) sont tous des éléments non-autonomes. Les SINEs sont composés de deux "boîtes" A et B qui sont deux sites de fixation à l'ARN polymérase III et d'une queue polyA à l'extrémité 3'. Ils mesurent de 100 à 500 pb et ne contiennent pas de terminateur de transcription pour l'ARN polymérase III [49, 163]. Cette ARN polIII est à l'origine de la transposition de la plupart

des éléments SINEs. Récemment, il a été montré que les éléments SINEs B1 pouvaient utiliser la machinerie de transposition des éléments LINEs [50]. Les SINEs sont divisés en deux sous-groupes : le groupe principal dérive des ARNt et le groupe minoritaire dérive de l'ARN 7SL [49] (Figure 1.9).

1.2.3 Les éléments transposables de classe II ou transposons

Dans cette section, nous décrivons la classe majoritaire des éléments transposables présents dans les génomes de plantes telles que le maïs ou *Arabidopsis thaliana* (génome étudié au cours de ce projet). Les hélitrons, éléments transposables principalement étudiés lors de cette thèse, sont des ETs de classe II.

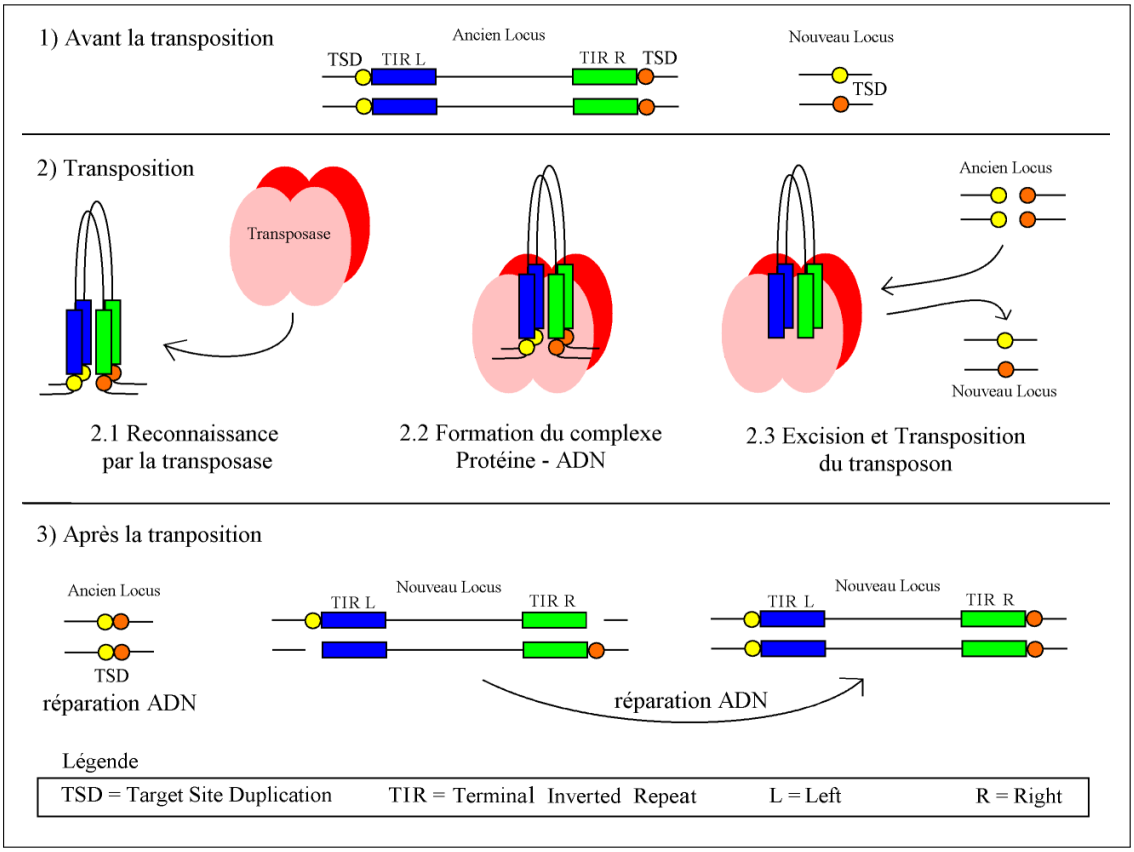


FIG. 1.10 – Mode général de transposition des éléments de classe II [39]. Après la transposition, le système de réparation de l'ADN par comparaison des deux chromatides sœurs, va recréer le transposon dans l'ancien locus. C'est cette réparation qui permet l'amplification de l'ET de classe II.

Cette classe est généralement caractérisée par la présence d'un TIR (Terminaison Inversée Répétée) aux extrémités de ces ETs. La structure palindromique des TIRs est reconnue par la transposase (la protéine de transposition) (Figure 1.10). La plupart des transposases de cette classe contiennent le motif catalytique DD34D/E. Contrairement aux rétrotransposons à LTR et rétroposons, les transposons à ADN (transposons au sens strict) ont été découverts aussi bien chez les procaryotes que chez les eucaryotes. La transposase coupe le transposon et le transpose dans un nouveau locus génomique. Ce nouveau

site est généralement identifiable par la présence du TSD. Le système de réparation de la cellule recopie le TSD simple brin, pour obtenir deux copies du TSD, caractéristique de cette transposition "couper-et-coller" (Figure 1.10).

1.2.3.1 Les transposons bactériens

On distingue également deux sous-catégories dans les transposons bactériens : la classe I regroupe tous les éléments transposables IS (Inserted Sequence) et les transposons composites formés d'IS, et la classe II regroupe les autres transposons bactériens, à l'exception des phages mutateurs.

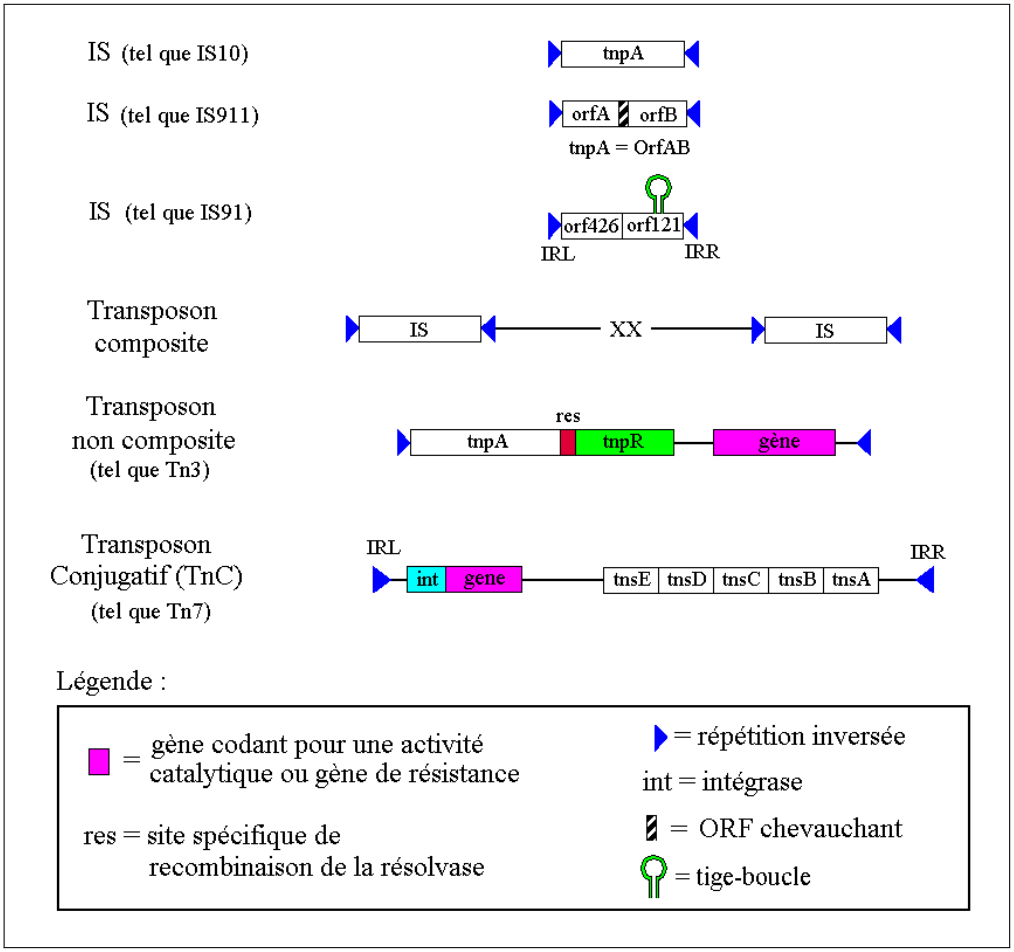


FIG. 1.11 – Structure et organisation des différents transposons bactériens [147, 146]

Les transposons IS sont les plus simples ETs connus : ils mesurent environ 2,5 kb et sont terminés par des IR (répétitions inversées) de 10 à 40 pb [39]. Ils codent pour une transposase (tnpA) (Figure 1.11). Certaines familles d'IS, par exemple l'IS911, ont une transposase codée par deux ORFs. Si deux IS sont assez proches l'un de l'autre, la transposase reconnaîtra les deux IS et déplacera l'ADN de l'hôte compris entre ces deux transposons : on parle alors de transposons composites (Ex Tn5) (Figure 1.11) [147].

Les transposons IS91 ont une structure plus complexe que les précédents IS avec deux ORFs, l'ORF426 contenant un gène de l'hélicase indispensable à la transposition

et l'ORF121 contenant un gène de la famille RPA (Replication Protein A) non indispensable à la réplication de l'IS [146]. IS91 est flanqué de deux courtes répétitions inversées appelées IRL et IRR et d'une tige-boucle subterminale (Figure 1.11) [146, 16]. Ce transposon s'insère au niveau de la séquence CTTG mais ne provoque pas sa duplication [52]. La répétition inversée IRL ne semble pas indispensable à la transposition d'IS91 [146] qui transpose via un mécanisme de rolling-circle analogue à certains plasmides et au phage $\Phi X174$ [16, 47]. IS91 possède beaucoup de caractéristiques similaires aux éléments transposables eucaryotes "hélitron" que nous étudions dans cette thèse.

La classe II des transposons bactériens est composée de deux grands types de transposons : les transposons non composites et les transposons conjugatifs. Les transposons non composites (Exemple : Tn3) sont composés de deux ORFs nécessaires à la transposition : la transposase (tnpA) et la résolvase (recombinase tnpR). Ils possèdent généralement un ORF codant pour un gène catalytique ou de résistance (RES sur la figure 1.11) et sont flanqués par des IR de 40 pb environ.

Les transposons conjugatifs (TnC) possèdent tous une intégrase, mais pas toujours d'IR (Figure 1.11). Les TnC sont des éléments chimériques qui possèdent les propriétés de plasmides conjugatifs et d'intégration des bactériophages. Ces éléments sont impliqués dans la dispersion de la résistance aux antibiotiques [147]. Les TnC existent dans la cellule sous deux formes distinctes : la forme linéaire, intégrée dans le génome (intégration similaire aux virus) et la forme circulaire dans le cytoplasme qui est la forme conjugative.

1.2.3.2 Les transposons eucaryotes

Cette classe regroupe sept grands types d'éléments transposables qui diffèrent par la structure de leurs TIRs et de leurs ORFs codant pour la transposase : les éléments de type bactérien, la superfamille TC1/mariner, la superfamille hAT, les éléments MuDR, les éléments de types Foldback, les Hélitrons et les Polintons.

Les éléments de type bactérien sont les éléments transposables les plus connus et les plus fréquents des transposons à ADN (Figure 1.12). L'élément le plus connu est sans doute l'élément P de la drosophile qui utilise un mécanisme de type "couper-et-coller" [39]. Sa découverte est associée à la mise en évidence de phénomènes dits de dysgénèse des hybrides chez la drosophile [103]. L'élément P mesure de 0,5 à 2,9 kb pour l'élément autonome [39], possède des TIRs de 31 pb indispensables à sa transposition et des répétitions inversées subterminales de 11 pb (Figure 1.12).

La superfamille Tc1/mariner est aussi une famille de transposons très étudiée et couvre la plupart des phyla eucaryotes allant des protozoaires aux mammifères. Les éléments Tc1/mariner sont parmis les éléments les plus simples des transposons eucaryotes. Ils mesurent de 1,3 à 2,4 kb [39]. La taille des TIRs est conservée à l'intérieur d'une même famille, mais varie de 20 à 40 nucléotides entre les membres de la superfamille [188]. Le TSD des Tc1/mariner est composé du dinucléotide TA [39, 188]. Les Tc1/mariner ne contiennent qu'un seul exon codant pour la transposase de 340 à 350 acides aminés, composée par la triade catalytique DD34D/E (Figure 1.12). [174, 7]. Par exemple, l'élément Hsmar1 mesure 1287 pb et a des TIRs de 30 pb [78].

La superfamille hAT contient deux sous-superfamilles : les éléments hAT et les éléments CACTA. En plus des TIRs, les éléments de ces deux superfamilles possèdent une séquence palindromique subterminale qui permet une régulation en *cis* de leur transposition [39, 208]. Les éléments hAT sont présents dans la plupart des eucaryotes tels que les levures,

les poissons et les humains [102]. Tous les éléments hAT présentent une structure similaire bordée par un TSD de 8 bp, mais la taille des TIRs est variable [102, 39]. La plupart des éléments hAT autonomes font plus de 4 kb. Les éléments Ac/Ds découverts par Barbara McClintock font partie de cette famille d'ETs. La superfamille CACTA présente des TSD de 3 bp et leurs gènes sont souvent constitués de plusieurs introns (Figure 1.12) [39]. Les TIRs ont aussi des séquences variables d'une famille à l'autre et mesurent de 10 à 28 pb. Ils sont constitués d'une région subterminale répétée de 10 à 20 pb [208].

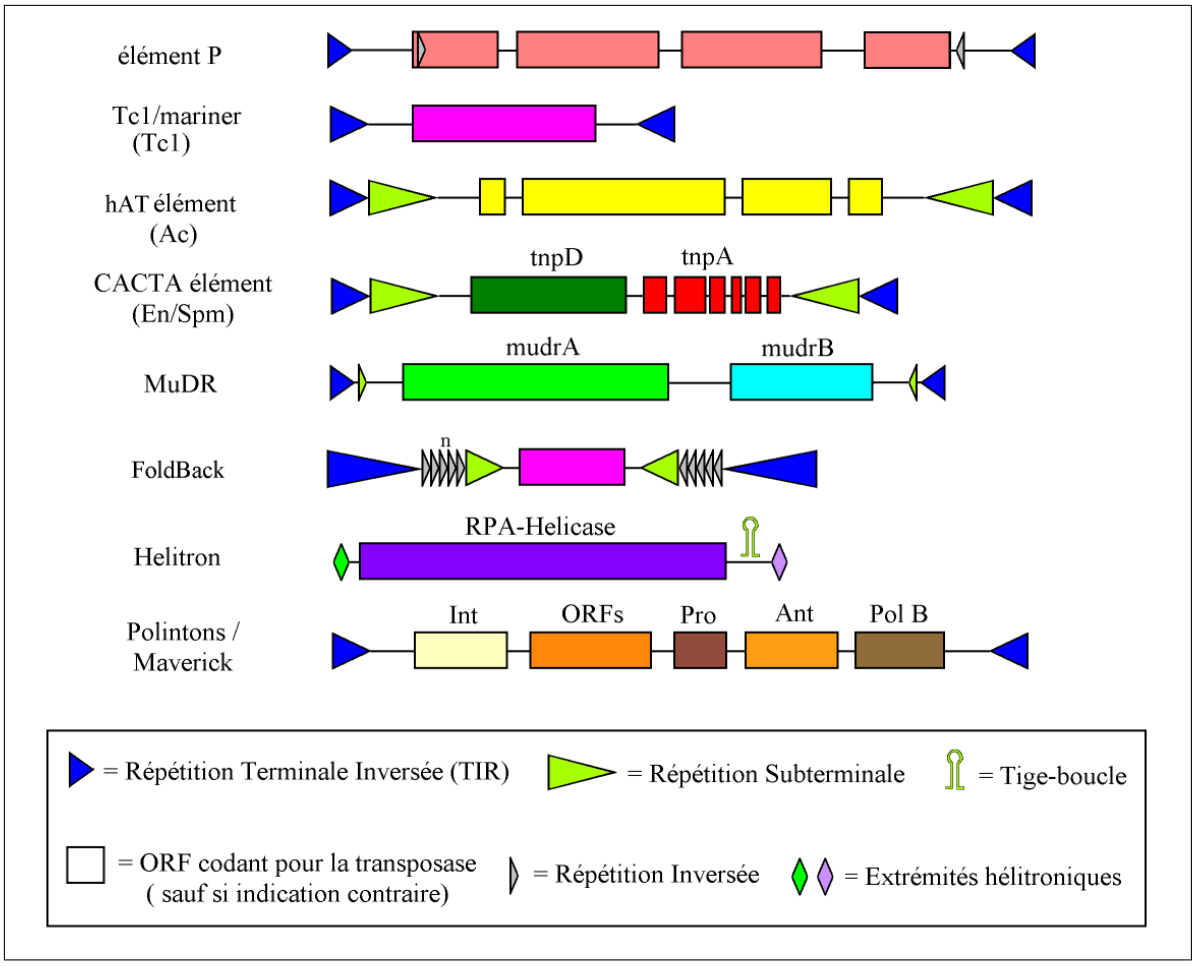


FIG. 1.12 – Structure et organisation des différents types de transposons eucaryotes [30].

Les éléments autonomes MuDR sont composés de deux gènes MudrA et MudrB (Figure 1.12) [51]. Les différentes familles sont bordées par des TIRs hautement conservés d'environ 200 bp. MudrA code pour une transposase (MURA) similaire aux protéines de transposition de certains transposons bactériens IS. MudrB code pour une protéine (MURB) non nécessaire à la transposition dont le rôle reste incertain [51, 187]. Les éléments non-autonomes ou MULEs (Mutator-like elements) sont mobilisés par la protéine MURA [187].

Les éléments de type Foldback [165] sont des transposons qui possèdent de très grandes inversions structurées en trois domaines : un domaine terminal, un domaine subterminal inversé répété de taille variable et un autre domaine répété un certain nombre de fois entre

les deux précédents [44] (Figure 1.12). Si leur structure rappelle les éléments de classe II, leur mode de transposition est encore aujourd'hui inconnu [39].

Les hélitrons sont des transposons caractérisés en 2001 par Kapitonov et Jurka [98]. Ils ont la particularité de ne posséder aucune répétition inversée ni de TSD. Les hélitrons sont le type d'éléments transposables étudiés au cours de cette thèse. Le chapitre 1.3 étant entièrement consacré à ce type hélitron, ils ne seront plus cités dans ce chapitre.

Les Polintons ou Maverick sont ses types d'éléments transposables de classe II découverts plus récemment [96]. Les Polintons sont caractérisés par des TIRs de plusieurs centaines de nucléotides et par la présence d'une DNA Polymérase B parmi leur ORFs. L'analyse phylogénétique semble indiquer que les Polintons sont proches des *Adenovirus*.

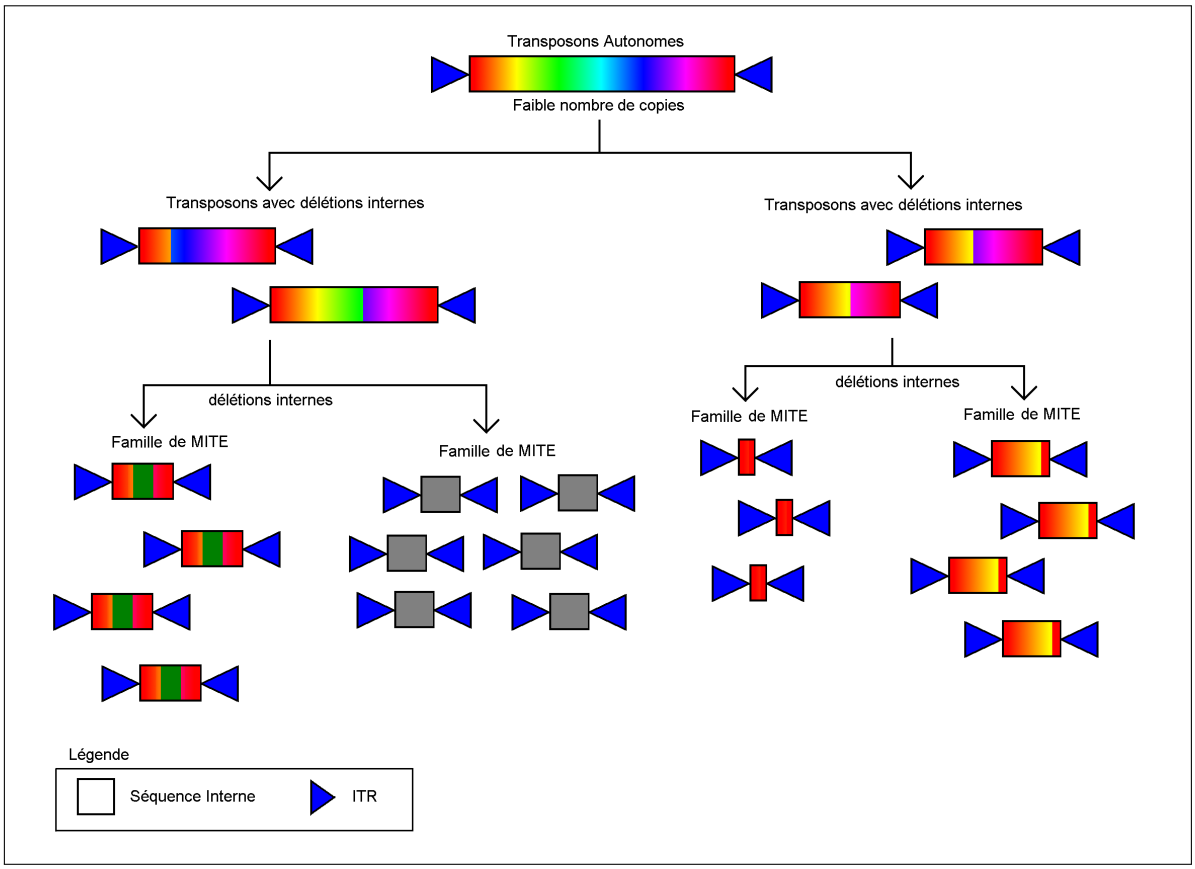


FIG. 1.13 – Création et amplification des MITEs à partir des transposons à ADN [39]. La palette de couleur marque les différentes portions de la séquence interne de l'élément autonome. Au cours de l'évolution, les copies de l'élément autonome subissent des délétions de séquences internes que l'on peut observer grâce à la délétion de certaines couleurs dans les éléments non-autonomes. Ces copies délétées subissent elles-mêmes des délétions/mutations pour former de courtes séquences appelées MITEs. Les MITEs possèdent généralement, en plus des TIRs, la séquence subterminale de l'élément autonome (3 familles sur 4 dans ce schéma). Les mutations de la séquence interne peuvent rendre impossible la reconnaissance de cette séquence par rapport à l'élément autonome (couleur grise et vert foncé sur le schéma).

Pour compléter ce panorama, il faut ajouter des sous-familles d'éléments non-autonomes (sans capacité propre de transposition), les MITEs (Miniature Inverted Transposable Elements), mis en évidence depuis quelques années dans la plupart des génomes [39,

68, 89]. Les MITEs sont issus de la délétion/insertion/mutation de la séquence interne de transposons autonomes de classe II (Figure 1.13) [68]. Si les MITEs, les plus connus sont les MITEs issus de la superfamille Tc1/mariner [159], les autres superfamilles de transposons peuvent aussi créer des MITEs, tels que le MITE Pack MuLE pour la superfamille Mutator [89] ou le MITE hATpin pour la superfamille hAT [148].

Les MITEs possèdent une grande homologie avec les TIRs et parfois avec la séquence subterminale des éléments autonomes (Figure 1.13). Ils présentent une structure secondaire stable (parfois plus stable que celle de l'élément autonome dont ils sont issus) et une très grande proportion de nucléotides A+T. Comme les SINEs et les LINEs, les MITEs sont plus nombreux dans les génomes que les éléments autonomes dont ils sont issus [39].

1.2.4 Rôles et importance des éléments transposables dans les génomes hôtes

En plus de leur rôle dans la plasticité structurale du génome, la plupart des ETs ont aussi un impact sur l'expression des gènes, la structure des régions promotrices, et même sur le fonctionnement cellulaire. Ils peuvent avoir un rôle positif, et dans ce cas le génome favorisera parfois leur invasion. Nous étudions d'abord la dynamique des ETs via leur proportion dans certains génomes. Nous présentons ensuite quelques exemples de rôles des ETs dans les génomes.

1.2.4.1 Proportions et distributions des éléments transposables dans les génomes

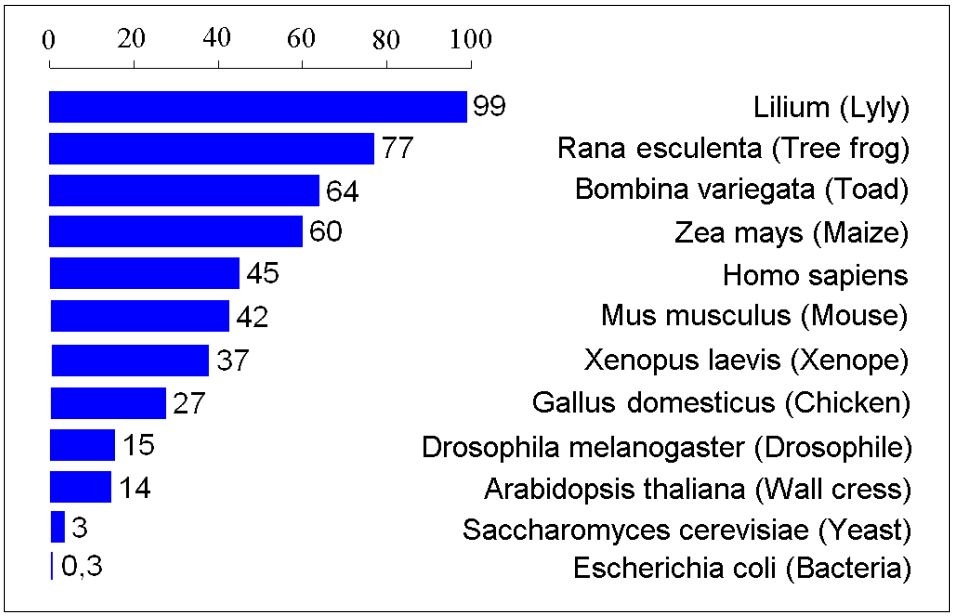


FIG. 1.14 – Proportion en pourcentage d'éléments transposables dans certains génomes modèles [106] © Christian Biémont.

Si les ETs sont présents dans la plupart des génomes séquencés, leur représentation dans ces génomes est variable [106, 104]. Les procaryotes, qui ont des génomes très com-

pacts et qui subissent des remaniements génomiques fréquents, possèdent une faible proportion d'ETs (Figure 1.14). Les eucaryotes possèdent une proportion plus variable d'éléments transposables. Chez les plantes, à l'exception du génome d'*Arabidopsis thaliana*, les génomes séquencés possèdent plus de 50 % d'ETs (Figure 1.14). Les amphibiens ont aussi une large proportion d'ETs dans leur génome. Une corrélation positive non généralisée (ne s'appliquant pas à tous les génomes) a été montrée entre la taille d'un génome et la proportion des ETs dans ceux-ci [104] : plus les ETs sont présents dans un génome et plus la taille du génome est importante.

Au sein d'un même génome, la capacité invasive et les proportions des différentes classes d'ETs sont aussi variables (Figure 1.15). Ainsi, la famille LINE est l'élément transposable majoritaire chez l'être humain, tandis que les rétrotransposons prédominent dans le génome du riz [83, 84].

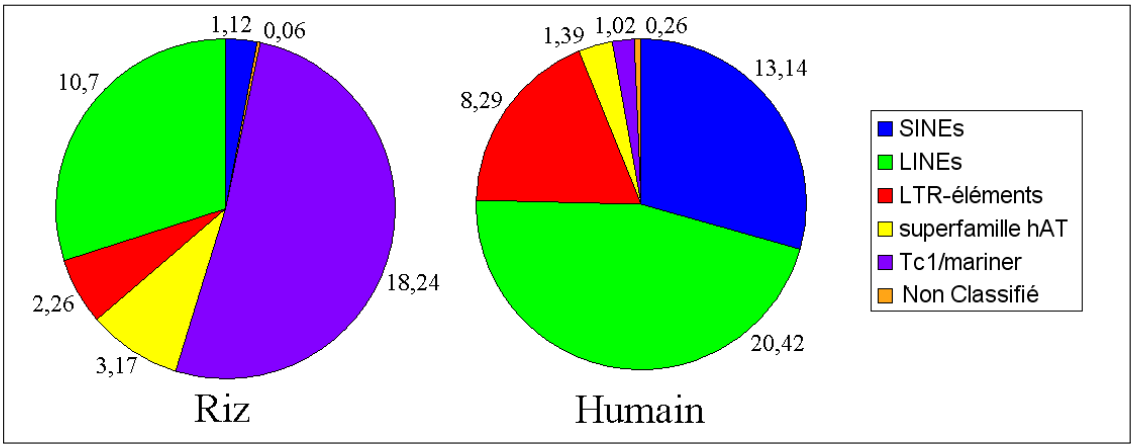


FIG. 1.15 – Proportion des différents types d'éléments transposables dans le génome du riz et de l'être humain [83, 84].

1.2.4.2 Rôles des éléments transposables dans le fonctionnement des gènes

Barbara McClintock a observé indirectement l'insertion de l'élément Ds dans le gène *waxy* du maïs [142]. Par la suite, les insertions des ETs dans les exons ont été très étudiées [39, 37, 105], comme les effets létaux ou les maladies qu'ils provoquent [28]. Il a été montré que leurs insertions ne correspondent pas réellement à des régions spécifiques du génome.

Même si leurs insertions n'ont pas lieu dans les exons, les ETs peuvent influencer les gènes situés à proximité. Les ETs peuvent aussi s'insérer dans les promoteurs, dans les introns ou les parties transcrites non traduites du gène.

L'insertion d'un élément transposable dans le promoteur peut modifier la transcription du gène [206]. Cette insertion peut inhiber les boîtes de régulation qui augmentent alors la transcription d'un gène [35] ou plus simplement éteindre le gène [91]. Mais l'insertion d'un élément transposable dans le promoteur d'un gène peut aussi donner un avantage au génome hôte, par exemple la création de nouveaux gènes par mutations [19, 20, 89].

Les effets des ETs sur l'activité des promoteurs ont été très étudiés chez les mammifères, en particulier dans le cas des rétrotransposons et des rétroposons [39]. Un élément transposable peut augmenter la transcription d'un gène situé à proximité (Figure 1.16).

Le rétrotransposon gypsy peut augmenter la transposition d'un gène en s'insérant dans son promoteur : la transcription du gène rapporteur de la luciférase est augmentée jusqu'à un facteur 8 [206] (Figure 1.16). De même, l'augmentation du nombre de transcrits d'un gène de drosophile marqué par un ET a été observé *in vivo* [110].

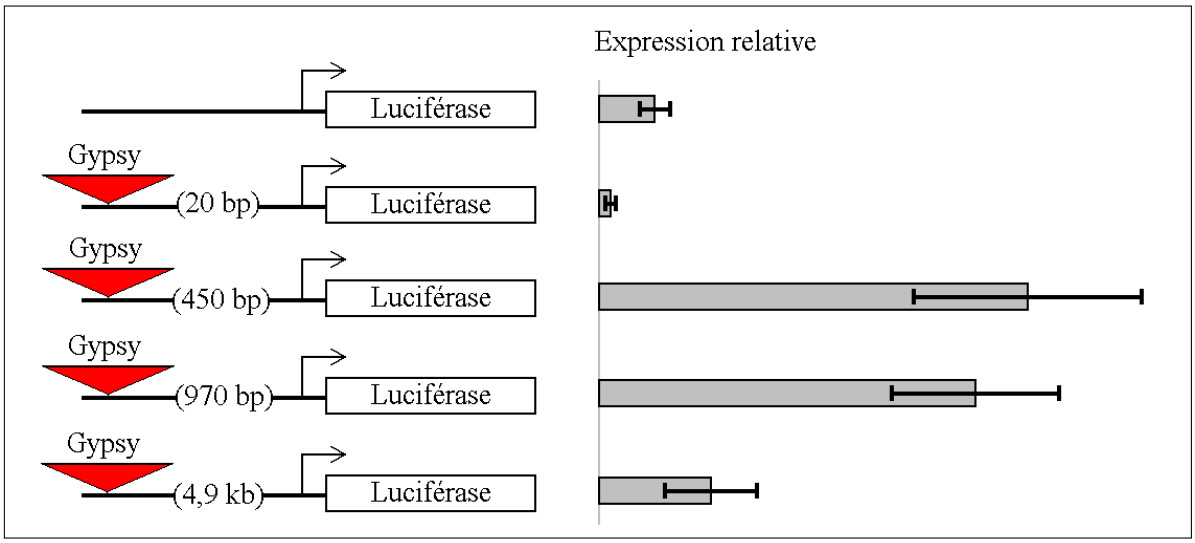


FIG. 1.16 – Activation du gène rapporteur de la luciférase par l'insertion du rétrotransposon de type *Gypsy* [206]. Si l'insertion de gypsy a lieu à une courte distance du codon START, elle provoque la diminution de la transcription du gène. Elle l'augmente pour une distance plus grande.

Les éléments transposables peuvent modifier la localisation tissulaire de la transcription d'un gène donné : chez l'humain et la souris, une dizaine de gènes orthologues ont une transcription tissulaire différente selon l'insertion ou non d'éléments transposables dans l'un des deux génomes [46].

D'autre part, les éléments non-autonomes peuvent apporter de nouveaux motifs de régulation contenus dans leurs séquences internes [164]. Chez l'humain, la séquence Alu contient plusieurs motifs de liaison aux facteurs de transcription [186] (Figure 1.17). Ces motifs vont modifier le profil d'expression du gène à proximité [39]. L'élément non-autonome peut s'insérer dans le promoteur et devenir une partie intégrante de celui-ci [14] ou devenir le nouveau promoteur du gène [56].

Même si les éléments transposables jouent un rôle dans la transcription des gènes en s'insérant dans leur région promotrice [39], à l'exception d'une étude de Thornburg sur les promoteurs humains [201], les analyses bioinformatiques des promoteurs retirent généralement la séquence de l'ET du promoteur pour l'étudier.

Un ET inséré dans un intron peut provoquer un épissage alternatif du gène [55] et l'insertion dans les régions transcrites non traduites peut créer un changement de stabilité de l'ARNm [53].

Si les ETs modifient la transcription des gènes du génome hôte, ils peuvent aussi modifier le nombre de gènes de ce génome. Lors de la transposition, si deux éléments transposables transposent au cours du même événement de transposition, l'ADN génomique situé entre ces deux transposons est aussi dupliqué dans un nouveau locus chromosomique [197, 39]. Le nombre de gènes dans le génome hôte peut aussi augmenter grâce à la capture

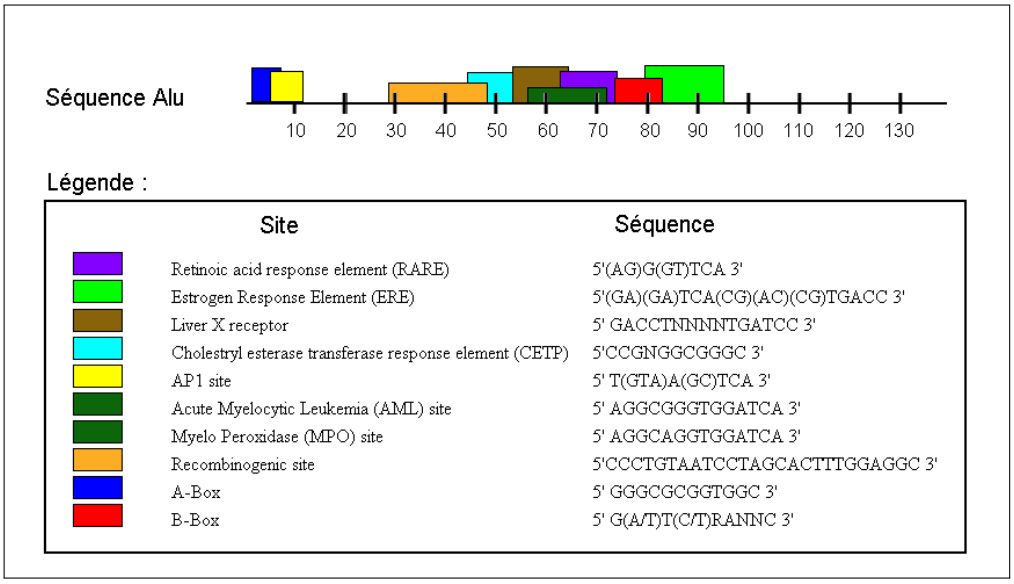


FIG. 1.17 – Motifs de régulation contenus dans une séquence consensus de 500 copies Alu [186].

de différents exons et à la création de nouveaux gènes paralogues [15].

1.2.4.3 Rôles des éléments transposables dans la création de nouveaux gènes

En plus de la capacité des ETs à contrôler l'expression d'un gène hôte, ils peuvent aussi se transformer en nouveau gène hôte. Cette transformation s'effectue après la perte de mobilité (perte des extrémités palindromiques et/ou des gènes de transposition dans le génome) et la mutation de la séquence interne de l'élément transposable. Certains éléments non-autonomes peuvent capturer des fragments de gènes différents [89] et peuvent être transcrits en un seul ARNm [149]. Ainsi, une récente étude a démontré que 5 % des exons du génome humain dérivent des rétrotransposons Alu [111].

1.2.4.4 Rôles des éléments transposables dans la structure du génome

Lors de la mitose, les séquences répétées peuvent être l'objet des recombinaisons non homologues (chapitre 1.1.1.2) [15]. Par exemple, la recombinaison non homologue d'éléments SINEs et la perte de séquence génomique ont provoqué l'apparition de cancers dans des lignées cellulaires humaines [39]. Ces recombinaisons n'entraînent pas que des pertes de séquences de génomes hôtes mais aussi des inversions de séquences [190]. Ainsi la recombinaison de deux éléments LINEs chez l'humain a provoqué l'inversion de 4 Mb du chromosome Y par rapport aux autres hominidés [39].

Certains éléments transposables s'insèrent préférentiellement dans l'hétérochromatine (portion de l'ADN du chromosome riche en ADN hyperméthylé contenant peu ou pas de gènes) [137, 190]. Ce processus joue un rôle essentiel dans les deux structures d'hétérochromatine du chromosome : le centromère et les télomères. Le centromère est la région d'hétérochromatine entre les deux bras du chromosome qui permet la ségrégation des deux chromosomes frères au cours la division cellulaire [39]. Les deux exemples les plus étudiés du rôle des ETs dans la structure centromère sont sans doute le rôle essentiel du transposon Tigger chez l'être humain et du transposon Pogo chez la drosophile [39].

Les télomères sont essentiellement composés de séquences répétées en tandem, mais chez la drosophile, ces répétitions sont remplacées par deux éléments transposables HetA et TART qui assurent le maintien des télomères au cours des divisions cellulaires par leur activité de transposition [39, 29, 138].

1.2.4.5 Autres rôles

Un autre rôle des éléments transposables a été largement étudié : la dysgénèse des hybrides chez la drosophile. Le croisement d'un mâle porteur de l'élément P et d'une femelle non porteuse provoque la transposition de l'élément P présent chez le zygote [39, 65, 171]. Ce phénomène se caractérise par des cassures chromosomiques, et des recombinaisons chez les mâles, entraînant la stérilité.

En situation de développement "normal", le génome hôte doit contrôler et inhiber la transposition des éléments transposables, par des mécanismes de contrôle épigénétique : l'hyperméthylation de l'ADN et les RNAi (RNA interference). La plupart des ETs sont hyperméthylés dans la plupart des stades de développement, pour éviter leur transposition [39, 112]. Ces deux mécanismes agissent à deux niveaux différents de la transcription. L'hyperméthylation de l'ADN empêche la transcription de l'ADN en ARNm grâce à l'ajout d'un résidu méthyl sur les cytosines [39]. Les RNAi empêchent la traduction de l'ARNm en protéine par appariement d'un ARN antisens sur l'ARNm, ce qui forme un ARN double brin reconnu et détruit par la machinerie cellulaire de l'hôte [39]. Le transposon CACTA chez *Arabidopsis thaliana* par exemple, est réprimé en présence de la protéine *DDM1* (protéine responsable de l'hyperméthylation des cytosines) et surexprimé si cette protéine est mutée [101]. Le rétrotransposon DIRS-1 chez *Dictyostelium* est réprimé par hyperméthylation de sa séquence mais aussi par RNAi si l'hyperméthylation n'est plus active [101]. Néanmoins, les éléments transposables peuvent apporter un avantage écologique et adaptatif pour l'organisme hôte lorsque l'hôte se retrouve en situation de stress environnemental. Une réponse adaptative de l'hôte pour tolérer ce stress peut être de relâcher son contrôle sur les éléments transposables et de provoquer des mutations grâce aux ETs. Cela lui permet de générer de la variabilité et d'offrir un plus large éventail de phénotypes compétitifs [74].

1.2.5 Détection *in silico* des éléments transposables

Les précédents exemples démontrent que les éléments transposables occupent une grande place dans la dynamique et le fonctionnement des génomes. Si depuis Barbara McClintock, les biologistes recherchent les ETs *in vivo*, la recherche d'éléments transposables *in silico* s'est développée après l'apparition des séquences génomiques (www.genomesonline.org) et de l'outil BLAST [3]. Pour rechercher les ETs, avant le début de cette thèse, deux types de méthodes existaient principalement : la recherche par similarité de séquences et la recherche *de novo* par filtres (Figure 1.18).

1.2.5.1 Détection par alignement de séquences

L'alignement d'une séquence requête contre une ou plusieurs autres séquences connues permet d'identifier cette séquence par similarité.

Il existe de nombreux logiciels qui utilise cette méthode. Parmi les plus connus nous pouvons citer RepeatMasker (www.repeatmasker.org) [191] et CENSOR (www.girinst.org-

/repbase/index.html) [95]. Ces deux programmes alignent la (les) séquence(s) requête(s) contre une bibliothèque de séquences. La bibliothèque de séquences d'ETs la plus connue est la bibliothèque de Repbase [94]. Elle regroupe les éléments transposables de plusieurs génomes et est mise à jour régulièrement. La dernière version de cette bibliothèque au moment de la rédaction de cette thèse est la version n°20061006. RepeatMasker est un ensemble de script PERL (www.perl.org) qui automatise le lancement d'un programme de comparaison de type BLAST (WU-BLAST ou Cross-Match). CENSOR fonctionne sur le même principe que RepeatMasker avec l'algorithme de comparaison de Smith et Waterman [192].

Cette méthode par alignement de séquences est la plus utilisée pour la détection d'éléments transposables [3, 13, 95, 94, 205]. Cette méthode a de plus l'avantage d'être, plus rapide et moins coûteuse en temps de calcul. Elle est principalement utilisée pour annoter les génomes séquencés tels que le riz ou le maïs, dont les séquences contiennent beaucoup d'éléments transposables [64, 92]. Néanmoins cette méthode a le désavantage d'être très conservative en ne recherchant que des éléments similaires à des éléments déjà connus.

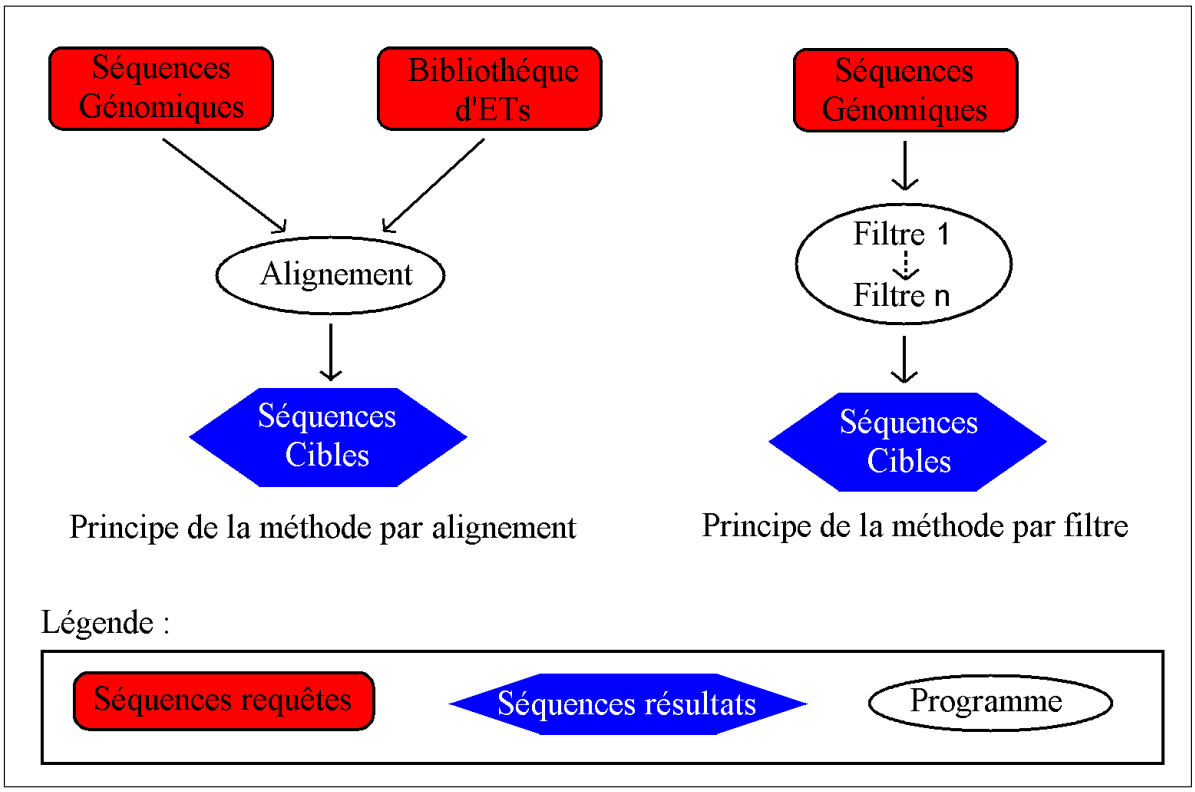


FIG. 1.18 – Les deux principales méthodes de détection des éléments transposables *in silico*.

1.2.5.2 Détection *de novo* des éléments transposables

Cette autre approche est basée sur la recherche d'éléments répétés à partir de la séquence source et de paramètres de détection tels que la taille de la répétition ou le nombre d'occurrences de cette répétition (Figure 1.18) [211]. Cette approche ne détecte pas seule-

ment les ETs mais aussi toutes les séquences répétées telles que les satellites ou les mini-satellites [189].

La plupart des logiciels de détection *de novo* réalisent un alignement de la séquence requête selon deux principales méthodes : l'alignement du génome contre lui-même et l'alignement du génome contre un génome phylogénétiquement proche. Parmi les logiciels qui utilisent l'alignement de la séquence contre elle-même, nous pouvons citer Recon [12], RepeatScout [167] et PILER [60]. Ils utilisent un dotplot (graphe matriciel) de l'alignement de la séquence contre elle-même puis en extraient les différents mots répétés à partir des différents paramètres de détection.

Parmi les logiciels alignant la séquence contre un génome phylogénétiquement proche, nous pouvons citer le logiciel de Caspi et Pachter [31] et le logiciel de Salse *et al.* [178]. Le premier programme a réalisé l'alignement de cinq génomes de drosophiles et le second a réalisé l'alignement d'Arabidopsis avec le génome du riz. Cette approche *de novo* a permis de découvrir de nouvelles familles d'éléments transposables, telle qu'une famille de *Gypsy* avec PILER [60] dans le génome de *Drosophila melanogaster*. Comme la précédente approche, elle ne permet pas de détecter les éléments qui ont une forte variation entre séquences et/ou un faible nombre de copies dans un génome.

1.2.5.3 Autres approches

Récemment, l'équipe de bioinformatique et génomique dirigée par Hadi Quesneville a créé un pipeline (suite de logiciels) dédié combinant les deux approches précédentes [170]. Cette stratégie a permis de ré-annoter les éléments transposables du génome de *Drosophila melanogaster*. Elle pose encore des problèmes de détection pour les éléments en faible nombre de copies dans le génome.

Trois autres logiciels basés sur des structures d'indexation de séquences : l'arbre des suffixes pour Reputer [116], le tableau des suffixes pour le logiciel de l'équipe de Rho *et al.* [173] et l'oracle des facteurs pour FORRepeats [124], permettent l'étude des éléments répétés dans un génome. Ces trois approches retrouvent parfaitement toutes les répétitions exactes quelle que soit leur taille, mais ont des difficultés de détection des éléments mutés notamment pour les éléments ayant subi des insertions-délétions. Le logiciel de l'équipe Rho est spécialisé dans la recherche de rétrotransposon à LTR. Il détecte deux répétitions maximales exactes (répétitions bornées par des contextes différents) d'au moins 40 pb et séparées d'un intervalle compris entre 1000 et 10000 pb, puis il aligne les séquences trouvées et sélectionne celles qui contiennent des domaines protéiques caractéristiques des rétrotransposons [173].

Nous terminons cette revue rapide par une méthode très différente des précédentes méthodes : LTR_STRUC [141]. LTR_STRUC recherche d'abord les deux LTRs grâce à leur similarité de séquences. A partir de ces fragments identiques de séquences, LTR_STRUC va aligner les deux LTRs et considérer qu'ils s'agit bien de deux LTR si les deux séquences alignées contiennent bien certains motifs spécifiques décrites dans le modèle LTR inscrit dans l'algorithme [141].

Si cette dernière méthode est restreinte aux rétrotransposons, cette approche par modèle d'un élément transposable basé sur ses caractéristiques biologiques est la démarche la plus proche de celle utilisée au cours de cette thèse et qui sera développée dans la partie : Matériels et Méthodes.

1.3 Hélitrons

Les hélitrons sont les éléments transposables qui font l'objet des études de cette thèse. Ce chapitre établit que ces éléments transposables sont très différents des autres ETs tant au niveau de leurs séquences et de leur structure qu'au niveau de leur mode de transposition et de leurs rôles dans les génomes. Ces différences ont suggéré de les classer dans certains articles comme des éléments de classe III (bien que les hélitrons soient des éléments de classe II du point de vue de leur mode de réplication ADN).

1.3.1 Contexte historique

Un nouvel ET détecté chez *Arabidopsis thaliana* en 1999 avait été nommé ATR0053 dans la base de données AtRepBase (nucleus.cshl.org/protarab/AtRepBase.htm). Cette même année, deux articles citaient ces éléments comme faisant partie de la famille AthEL-1.4 [195] et de la famille AtREP [97]. Ces éléments répétés ne semblaient avoir ni structure particulière à leurs extrémités ni séquence interne consensus. A cause de ce manque de caractérisation, Thomas Bureau, croyant les découvrir, les nommera Basho [123] en 2000. Ce n'est qu'en 2001 que Vladimir Kapitonov et Jerzy Jurka donneront le nom hélitron et les caractéristiques de cet élément transposable [98].

La découverte d'un élément transposable de type hélitron dans le promoteur du gène ADC1 (Arginine DéCarboxylase 1) chez *Arabidopsis thaliana* [62], qui est un gène essentiel de la biosynthèse des polyamines, a conduit l'équipe Expression génétique et adaptation de l'UMR 6553 EcoBio et l'équipe Symbiose de l'INRIA de Rennes à proposer cette étude de bioinformatique des hélitrons chez *Arabidopsis thaliana*.

1.3.2 Définition

1.3.2.1 Caractéristiques structurales

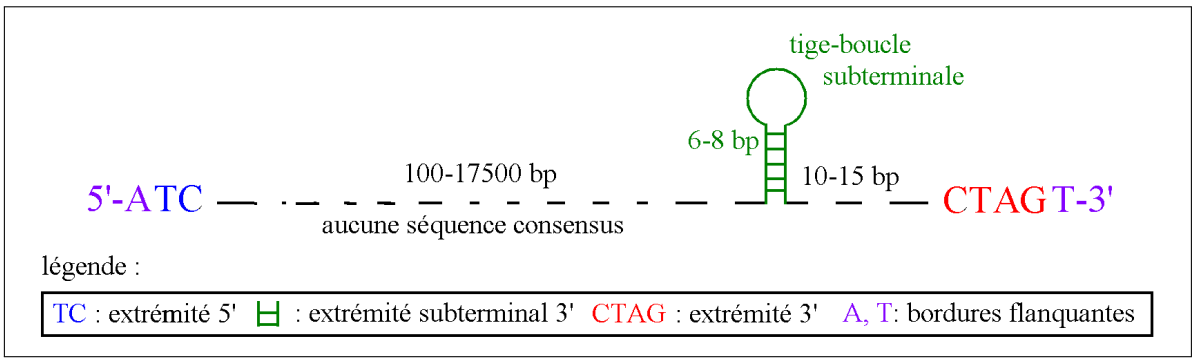


FIG. 1.19 – Modèle de l'hélitron proposé par Kapitonov et Jurka [98].

Un hélitron est un élément transposable de classe II caractérisé par le dinucléotide TC à l'extrémité 5' et CTAG avec une tige-boucle (hairpin) subterminale à l'extrémité 3' (Figure 1.19). L'hairpin a une composition riche en G+C, mais n'a pas de séquence ni de taille consensus (Figure 1.19). La séquence interne des hélitrons est très variable

pour les éléments non-autonomes. Les éléments autonomes présentent plusieurs ORFs codant une hélicase et une protéine RPA-like [98]. Les hélitrons ne sont pas bordés par des répétitions directes mais seulement entourés du nucléotide A en 5' et T en 3'. Chez *Arabidopsis thaliana*, certains hélitrons non-autonomes s'appellent AtREP (*Arabidopsis thaliana* REPetitive element) [97].

1.3.2.2 Mode de transposition putatif des hélitrons

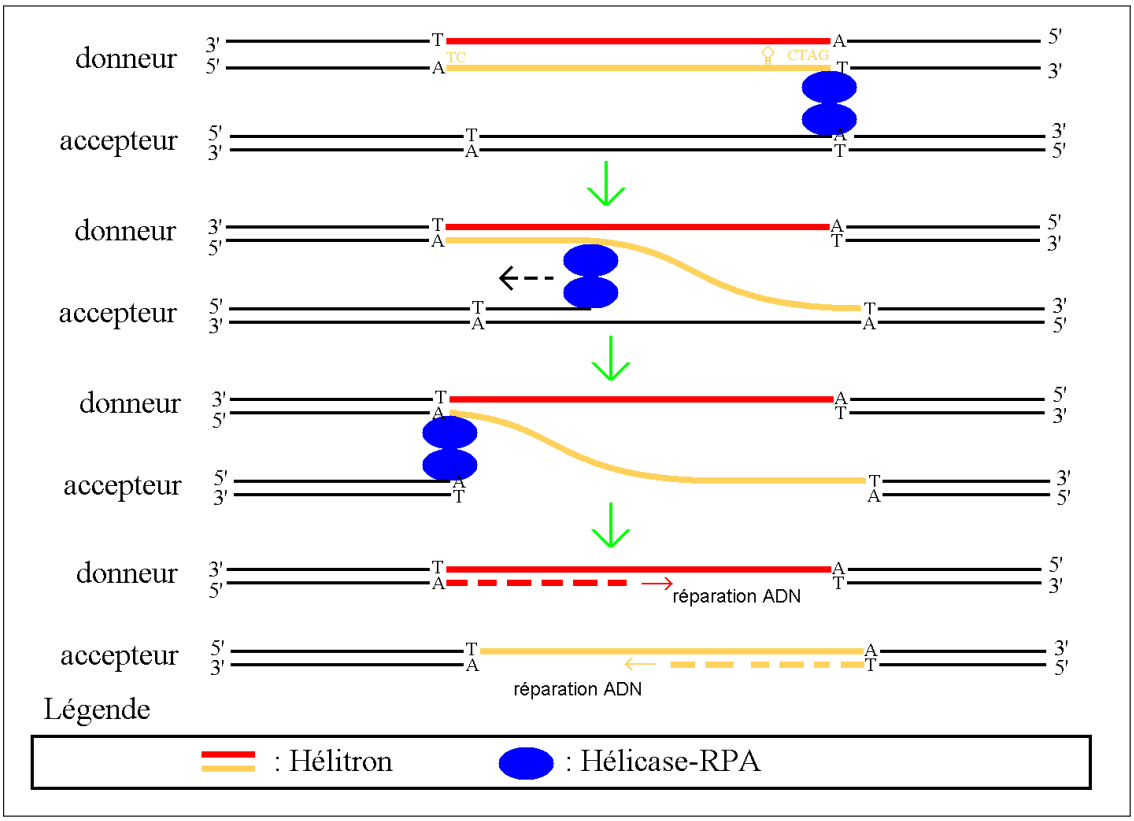


FIG. 1.20 – Modèle putatif de la transposition d'un hélitron [16, 70]. Un seul brin (couleur orange) est transposé par le mécanisme de transposition rolling-circle. La cellule répare l'ADN simple brin.

La structure des hélitrons et de ses protéines de transposition rappelle le mode de transposition par cercle roulant (ou rolling-circle) [70, 98] utilisé par deux autres types d'éléments mobiles : les Géminivirus [77, 122] et les IS91 des procaryotes [16, 52, 146, 145]. On pense actuellement que lors d'une transposition d'un hélitron, la transposase (nom donné au complexe protéique hélicase-RPA [70]) coupe l'extrémité 3' du brin donneur grâce à la reconnaissance de l'hairpin et coupe le brin accepteur (Figure 1.20) [70, 145]. Puis le complexe hélicase-RPA (Replication Protein A) va lier l'extrémité 3' de l'hélitron sur le brin accepteur. Elle coupe ensuite l'extrémité 5' de l'hélitron sur le brin donneur et l'attache sur le brin accepteur (Figure 1.20) [16]. Ce mode de transposition permet à l'hélitron transposé d'être dans le même sens que l'hélitron originel ou dans le sens inverse [16]. Si ce mode de transposition n'a pas été expérimentalement démontré à ce jour, l'activité de transposition d'un hélitron a été prouvée dans le génome du maïs [119, 120].

1.3.2.3 Répartition des hélitrons dans les génomes

Initialement découverts dans le génome modèle *Arabidopsis thaliana* [195], les hélitrons ont été ensuite retrouvés dans le génome d'*Oryza sativa* (riz) et de *Caenorhabditis elegans* (ver) [98]. De nouveaux hélitrons ont été trouvés dans deux génomes de poisson : *Danio rerio* et *Sphoeroides nephelus*, ainsi que dans les génomes des levures *Phanerochaete chrysosporium* [166] et *Aspergillus nidulans* [42]. Ils ont aussi été découverts dans deux génomes entièrement séquencés d'insectes : *Drosophila melanogaster* et *Anopheles gambiae* [99, 166]. Enfin de nombreuses études montrent la présence d'hélitrons dans le génome de *Zea mays* (maïs) [23, 58, 75, 119, 121, 120, 149]. Dernièrement, les hélitrons ont été découvert dans deux nouveaux génomes : la fleur *Ipomoea tricolor* [32] et la chauve-souris *Myotis lucifugus* [169]. La découverte des hélitrons dans ces génomes a été réalisée par une méthode d'alignement (généralement BLAST [3]) du génome contre lui-même. Néanmoins, la faiblesse des connaissances sur le modèle hélitronique peut laisser penser que les hélitrons sont présents dans d'autres génomes eucaryotes séquencés, mais qu'ils n'ont pas encore été détectés.

1.3.3 Relation Hôte - Hélitron

Si les hélitrons sont présents dans plusieurs génomes eucaryotes, leur activité et leurs rôles fonctionnels n'ont été étudiés que chez le maïs. En plus de leur rôle mutateur dû à la transposition et de leur effet sur l'épissage alternatif [121], deux effets ont été étudiés : la réorganisation du génome et la création de nouveaux gènes.

1.3.3.1 Effet des hélitrons dans la création de gènes

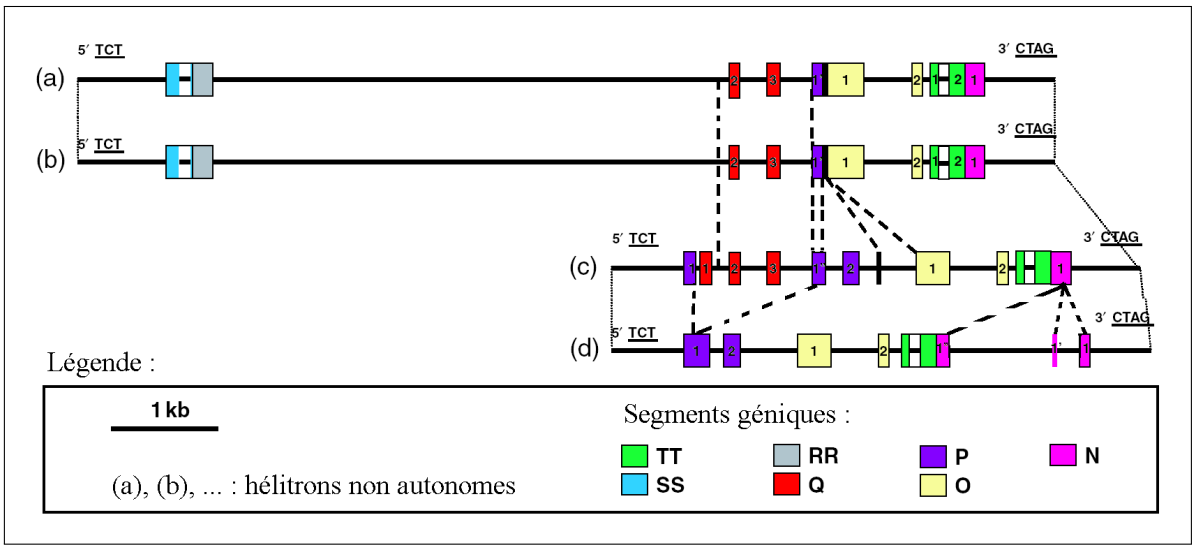


FIG. 1.21 – Exemple d'impact des hélitrons non-autonomes qui ont capturé des segments géniques chez le maïs [23].

Chez le maïs, certains hélitrons non-autonomes ont capturé des séquences génomiques de l'hôte tels que des pseudogènes [75] ou des exons [121] provenant de différents gènes

(Figure 1.21) [23]. De plus, ces exons montrent une activité de transcription qui démontre le rôle des hélitrons dans la création d'un gène par l'assemblage d'exons différents (Figure 1.21) [15, 23].

Cette capture de séquence hôte est due au mode de transposition par rolling-circle de l'hélitron qui permet la capture de séquences génomiques à proximité du brin donneur [68]. Cette capture de gènes démontre le rôle évolutif que peuvent jouer les hélitrons au sein des génomes. La capture puis la transposition de l'hélitron entraînent la duplication du gène capturé [149]. Le gène dupliqué peut alors muter pour donner de nouvelles fonctions ou permettre à l'organisme une expression différentielle des deux gènes [149].

1.3.3.2 Hélitron et structure du génome

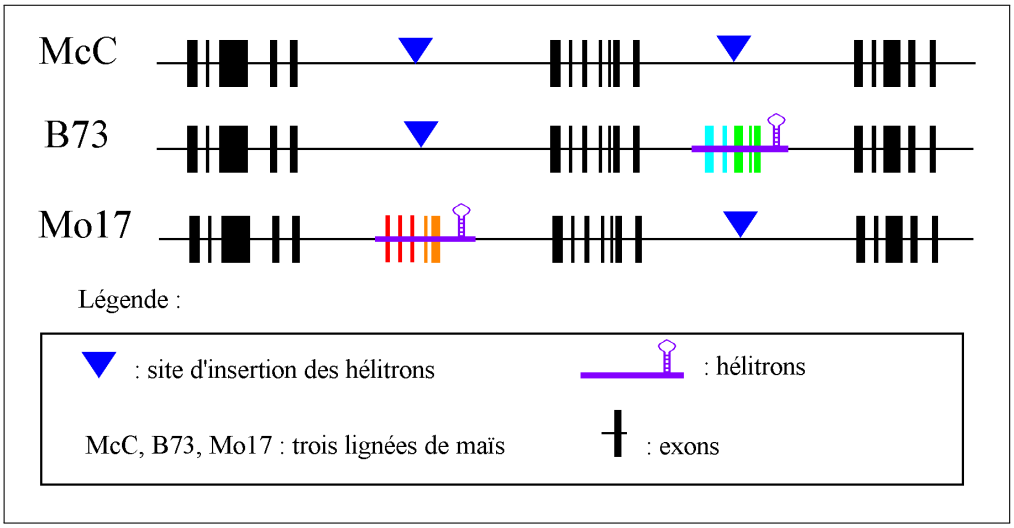


FIG. 1.22 – Exemple de perte de colinéarité entre trois lignées de maïs [120]. L'insertion d'hélitron avec des exons dans sa séquence interne entraîne une perte de colinéarité des gènes entre les lignées. Les triangles bleus indiquent les sites d'insertion possibles des hélitrons.

Chez le maïs, les hélitrons transposent, activent et transportent dans leur séquence interne des exons, des pseudogènes et même des portions non codantes du génome [75, 119, 121, 149]. En plus du déplacement de séquences génomiques capturées, deux hélitrons peuvent transposer ensemble et entraîner la transposition de toute la séquence de l'hôte présente entre eux [120]. Cette activité entraîne de larges mutations et remaniements chromosomiques entre les lignées de maïs (Figure 1.22) : par exemple il y a 20 % de variation de colinéarité de gènes entre les lignées *B73* et *Mo17* (sur l'ensemble du génome) [149].

Tous ces résultats montrent l'importance des hélitrons sur l'évolution du génome hôte, d'où l'importance de comprendre leur fonctionnement et leur mécanisme de transposition.

1.4 But de la thèse

Au début de ce travail en 2004, seuls deux articles décrivaient les hélitrons [98, 70] (Figure 1.23). Il a fallu attendre deux années pour démontrer leur activité de transposition. Leur premier rôle fonctionnel a été publié en 2005 concernant leur implication dans la création de nouveaux gènes à partir de la capture de différents exons [23]. Néanmoins, ces résultats ont été obtenus sur des lignées agro-industrielles de maïs et ne sont donc pas dans le domaine public, ce qui ne permet pas de réaliser de nouvelles analyses. Ainsi, les connaissances sur les hélitrons sont en fait essentiellement limitées à la librairie d'éléments transposables (séquence au format FASTA) contenue dans la base de données Repbase [94]. Nous avons choisi de travailler sur le génome modèle *Arabidopsis thaliana*, car ce génome est séquencé, annoté et parce qu'il possède de nombreuses sources d'informations pour une approche bioinformatique. La figure 1.24 résume l'ensemble des questions qui ont été abordées dans ce contexte.

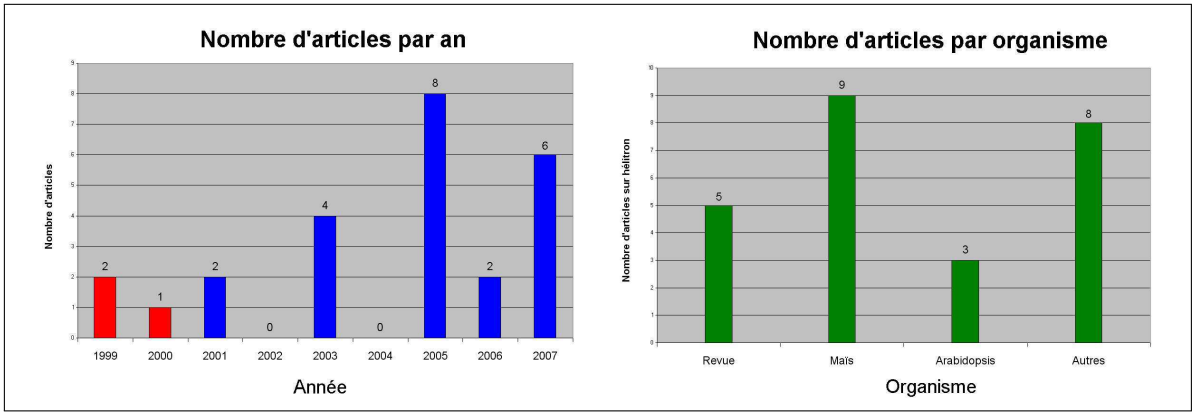


FIG. 1.23 – Nombre d'articles parus sur les hélitrons depuis leur découverte et nombre d'articles par génome. Sur l'histogramme de gauche, la couleur rouge représente les articles avant la définition des hélitrons donnée par Kapitonov et Jurka [98], la couleur bleue représente les articles après cette définition.

1.4.1 Problèmes posés pour la détection *in silico* des hélitrons

Le manque de structure aux extrémités et l'absence de séquence consensus à l'intérieur des hélitrons [98] rendent très difficile leur détection *in silico*. Les méthodes classiques de détection *de novo* tel que Recon [12] et RepeatScout [167] ne peuvent pas détecter ces éléments transposables qui ont des séquences internes très variables ou présentes en faible nombre de copies telles que la famille hélitron2 chez *Arabidopsis thaliana* [98]. La variabilité des hélitrons empêche aussi leur détection dans de nouveaux génomes : ainsi, la découverte d'hélitrons chez *C. elegans* n'a été réalisée que deux ans après celle d'Arabidopsis [99] alors que le génome du ver était accessible avant celui de la plante [63, 197]. Actuellement, seules les approches combinées d'Hadi Quesneville [170] et l'alignement génome contre génome permettent de détecter partiellement ces éléments.

Une première question est apparue sur la séquence des hélitrons non-autonomes comme AtREP3 : comment retrouver toutes les séquences des hélitrons avec si peu de motifs

caractéristiques (Figure 1.24) ? Si les hélitrons autonomes possèdent des ORFs et des domaines protéiques "facilement" identifiables, ces ORFs n'existent pas dans les hélitrons non-autonomes or ceux-ci représentent la majorité des hélitrons dans le génome d'*Arabidopsis thaliana*. De plus, à l'exception de ce génome, aucun autre hélitron non-autonome sans ORF (ORF de transposition ou autre ORF) n'a été découvert et décrit dans d'autres génomes. Il était important dans un premier temps de trouver une nouvelle méthode de détection des hélitrons, et des éléments transposables en général. La découverte de nouveaux motifs caractéristiques des hélitrons était primordiale pour récupérer l'ensemble des séquences des hélitrons dans le génome d'*Arabidopsis thaliana* et les analyser ensuite dans ce projet. Nous avons donc recherché les motifs caractéristiques des hélitrons et détecté l'ensemble des hélitrons présentant ce motif dans le génome d'*Arabidopsis thaliana*. Cette recherche est abordée dans les deux premiers chapitres des résultats de cette thèse.

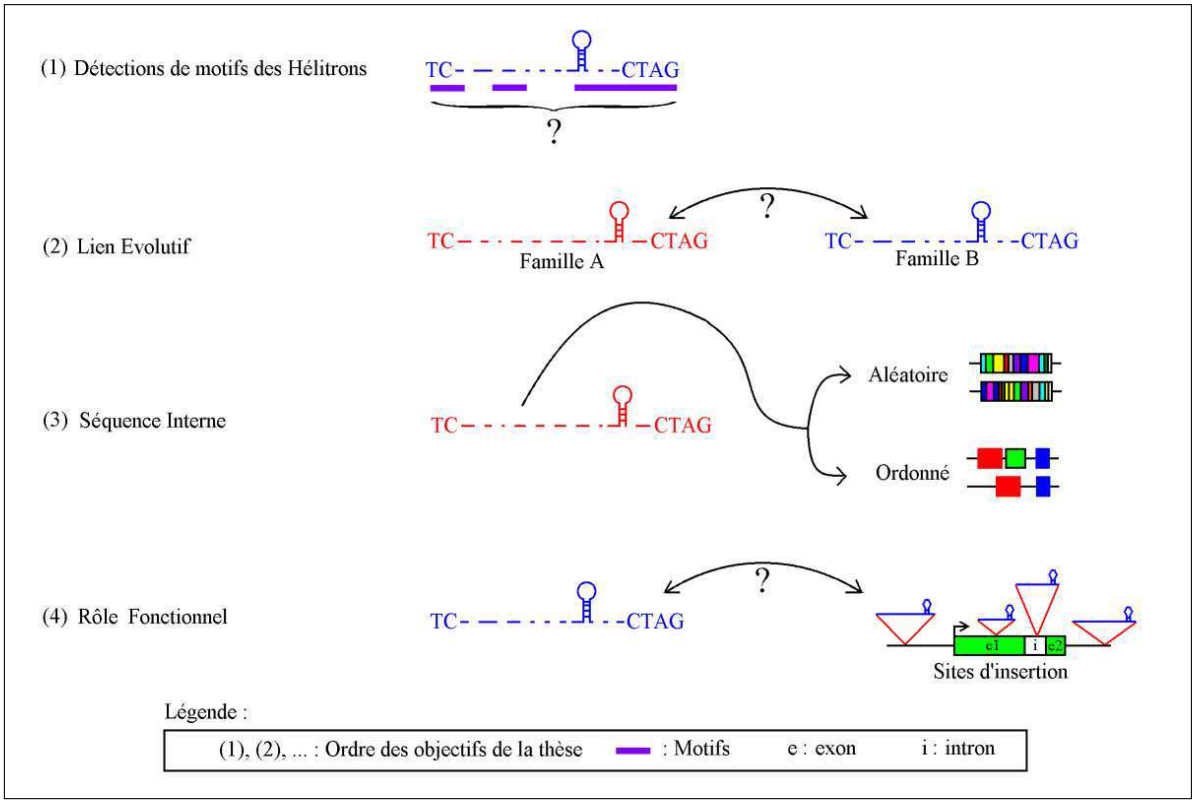


FIG. 1.24 – Principaux objectifs de ce projet : Modélisation, classification, organisation de la séquence interne des hélitrons et rôle fonctionnel au sein du génome d'*Arabidopsis thaliana*.

1.4.2 Interaction des hélitrons avec le génome d'*Arabidopsis thaliana*

Une précédente étude montrait que 1 % de ce génome est composé d'hélitrons, pour la plupart non-autonomes [98]. Mais la variabilité des séquences internes empêchait toute analyse phylogénétique de cette famille d'éléments transposables [98]. Aucune étude n'a-

vait répondu alors à cette première question : comment des hélitrons non-autonomes ont pu envahir ce génome (Figure 1.24) ? Il était donc important de comprendre les mécanismes de transposition des hélitrons pour retracer l'évolution des hélitrons. Cette interaction des différentes familles d'hélitrons autonomes et non-autonomes est soulevée dans le troisième chapitre des résultats.

1.4.3 Variabilité des hélitrons

Cette question amène directement deux autres questions sur la variabilité de la séquence interne des hélitrons : quel phénomène biologique est à l'origine de la variabilité des séquences internes des hélitrons (Figure 1.24) ? Quel est l'effet de cette séquence interne sur le génome hôte ? En effet, de précédentes études montraient la grande variabilité de la séquence interne des hélitrons autonomes [98] mais aucun phénomène biologique n'expliquait cette diversité sinon la capture de gènes [15]. Aucune étude non plus ne montrait de capture de gènes par l'hélitron chez *Arabidopsis thaliana*, ni leur effet sur la dynamique de ce génome. Il était donc important d'appréhender les mécanismes possibles qui modifient la séquence interne de ces éléments. Ces deux questions sont traitées dans les quatrième et cinquième chapitres des résultats de cette thèse.

1.4.4 Effet des hélitrons sur le génome d'*Arabidopsis thaliana*

La découverte d'un hélitron par l'équipe Expression génétique et adaptation apporte enfin ces dernières questions sur l'interaction des hélitrons sur le génome hôte : apportent-ils un effet bénéfique au génome hôte ? Capturent-ils des gènes comme chez le maïs et/ou transportent-ils des éléments de régulation comme chez la levure ? Ces questions sont abordées dans le dernier chapitre des résultats.

L'ensemble de ces interrogations montrent que les hélitrons peuvent être considérés comme un élément modèle dans la dynamique et la variation du génome hôte grâce à leur capture de séquences hôtes, à la création de nouveaux gènes [149], à leur grande variabilité de séquence interne qui peut entraîner la création de nouvelles séquences utiles au génome, et au fait de leur forte proportion dans ces génomes [98].

Matériels et Méthodes

Ce chapitre permet de présenter dans un premier temps Arabidopsis et son génome, les données et les choix des sources de données utilisées au cours de cette thèse. Nous parlerons ensuite des méthodes d'analyses de ces données référencées dans la littérature. Nous commençons par une brève description de l'organisme étudié.

2.1 Le génome d'*Arabidopsis thaliana*

Le travail de cette thèse a été réalisé sur le génome de la plante modèle *Arabidopsis thaliana* qui était déjà étudié dans l'équipe laboratoire "Expression génétique et adaptation" (UMR EcoBio).

FIG. 2.1 – Photos de la plante *Arabidopsis thaliana* à différents stades de son développement (www-ijpb.versailles.inra.fr/fr/sgap/equipes/cyto/arabido.htm).

Arabidopsis thaliana ou "Arabette de Thal" ou encore "Arabette des dames" fait partie des angiospermes (plante à fleur) appartenant à la famille des *Brassicaceae*. Cette plante est devenue un organisme modèle du fait d'un certain nombre de caractéristiques :

- Arabidopsis est une petite plante et nécessite peu de place pour se développer (Figure 2.1).
- cinq à huit semaines suffisent depuis la germination de la graine jusqu'à la production de graine par l'adulte.
- les études génétiques sont facilitées par une production abondante de graines par plante (environ 10000).
- Arabidopsis est une plante auto-pollinisatrice qui permet d'obtenir simplement des individus homozygotes (individus ayant les deux allèles identiques pour le même gène) et d'étudier les mutations récessives.
- des plants transgéniques sont facilement créés avec la bactérie *Agrobacterium tumefaciens* utilisée comme vecteur d'insertion de gènes.

Arabidopsis thaliana est le premier génome de plante entièrement séquencé [197] en 2000. C'est l'un des plus petits génomes de plante avec 115 409 949 nucléotides répartis sur 5 chromosomes. *Arabidopsis thaliana* possède aussi un génome mitochondrial de 350000 bp et un génome plastidial de 150000 bp [197]. L'annotation du génome (TAIR version 6.0 du 11 novembre 2005 www.arabidopsis.org) mentionne 26751 gènes, 3818 pseudogènes et 838 ARN non codants.

Le projet MASC NSF (Multinational Arabidopsis Steering Committee National Science Fundation) 2010 Project Arabidopsis (www.arabidopsis.org/info/2010_projects/) a été lancé en 2002 [8]. Ce projet de génomique, transcriptomique, protéomique et bioinformatique veut élucider la fonction de tous les gènes appartenant à certains réseaux présents dans le génome d'*Arabidopsis thaliana*, avec des études plus ciblées sur l'expression et le réseau des gènes liés au MAPK (Mitogen-activated protein kinase), les signaux d'épissage présents dans les préARNm ou encore les gènes liés au développement de la graine. Le projet consiste aussi à développer des outils dédiés de recherche et le stockage des données [8].

2.2 Bases de données utilisées

Nous avons répertorié l'ensemble des données utilisables (téléchargeables ou non). Nous avons privilégié les bases de données téléchargeables. Ces données sont réparties en trois catégories : les données sur le génome d'*Arabidopsis thaliana* (séquences et annotations), les expérimentations sur les hélitrons et les données d'expression de ce génome (essentiellement des données de microarrays).

2.2.1 Bases de données sur Arabidopsis

De nombreuses bases de données ont été créées pour stocker toutes les données concernant le génome d'*Arabidopsis thaliana* (Figure 2.2). La base de données la plus connue est sans doute la base TAIR (The Arabidopsis Information Ressource www.arabidopsis.org) [172]. Cette base contient, en plus du génome complet, les annotations du génome réalisées par le TIGR (The Institute for Genomic Research), Arabidopsis thaliana Database (www.tigr.org/tdb/e2k1/ath1/), l'ensemble des articles scientifiques concernant les gènes, les profils d'expression des gènes provenant de différentes bases de données et beaucoup d'autres données et fonctionnalités. De plus, TAIR vérifie la qualité des données. En effet, une étude préliminaire avait montrée des erreurs dans les positions des annotations du TIGR. Nous avons donc choisi cette base de données pour télécharger le génome et ses annotations.

Thème de la base	Nom de la base de données	Site Web	Références
Généraliste	The Arabidopsis Information Ressource (TAIR)	www.arabidopsis.org/	Rhee 2003
Généraliste	TIGR Arabidopsis thaliana Database	www.tigr.org/tdb/e2k1/ath1/	
Généraliste	RIKEN Arabidopsis Genome Encyclopedia (RARGE)	rarge.gsc.riken.go.jp/index.html	Sakurai 2005
Généraliste	Munich Information center for Protein Sequence (MIPS)	mips.gsf.de/proj/thal/db/	Mewes 2006
microARN	Arabidopsis Small RNA Project (ASRP)	asrp.cgrb.oregonstate.edu/	Gustafson 2005
Développement	The Arabidopsis SeedGenes Project	www.seedgenes.org	Tzafrir 2003
Métobolisme	The Arabidopsis Lipid Gene Database	www.plantbiology.msu.edu/lipids/genesurvey/	Mekhedov 2000
Elément Transposable	Arabidopsis thaliana Genome Analysis	nucleus.cshl.org/protarab/	
Elément Transposable	Repbase	www.girinst.org/repbase/index.html	Jurka 2005
Elément Transposable	Arabidopsis Transposable Element Database	www.tebureau.mcgill.ca/clonebase/main.html	Le 2000
Elément Transposable	Arabidopsis thaliana Integrated Database	atidb.org/	Pan 2003
Expression de Gène	Arabidopsis Functional Genomics Network (AFGN)	www.uni-tuebingen.de/plantphys/AFGN/	Schmid 2005
Expression de Gène	Complete Arabidopsis Transcriptome MicroArray (CATMA)	www.catma.org/Database/index.html	Crowe 2003
Expression de Gène	Nottingham Arabidopsis Stock Centre's micro Arrays (NASCArrays)	affymetrix.arabidopsis.info/	Craigon 2004
Expression de Gène	Gene Expression Omnibus (GEO)	www.ncbi.nlm.nih.gov/geo/	Barrett 2005
Expression de Gène	Genevestigator	www.genevestigator.ethz.ch/at/	Zimmermann 2004
Expression de Gène	The Arabidopsis Gene Expression Database (AREX)	www.arexdb.org/index.jsp	
Facteur de transcription	RIKEN Arabidopsis Transcription Factor database (RARTF))	rarge.gsc.riken.jp/rartf/	Iida 2005
Facteur de transcription	Database of Arabidopsis Transcription Factors (DATF)	datf.cbi.pku.edu.cn/index.php	Guo 2005
Facteur de transcription	Plant cis-acting regulatory elements (PlantCARE)	bioinformatics.psb.ugent.be/webtools/plantcare/html/	Lescot 2002
Facteur de transcription	Arabidopsis Gene Regulatory Information Server (AGRIS)	arabidopsis.med.ohio-state.edu/	Davuluri 2003

FIG. 2.2 – Les principales bases de données comportant des données sur *Arabidopsis thaliana*.

2.2.2 Bases de données d'éléments transposables

De nombreuses bases de données spécialisées sur les ETs existent sur le Web. La plupart de ces bases de données sont spécialisées sur un organisme précis, comme par exemple

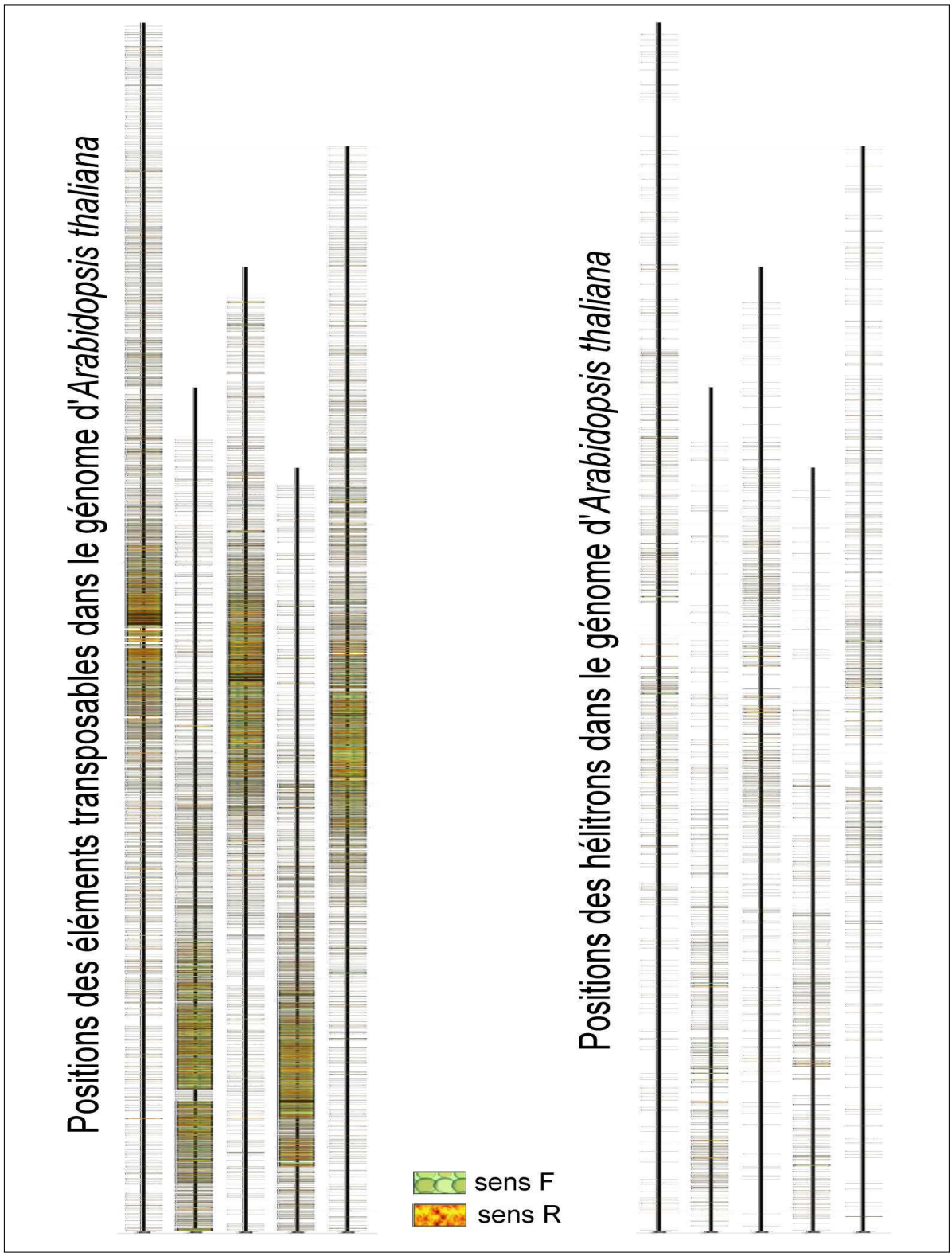


FIG. 2.3 – Positions des ETs et de hélitrons présents dans le génome d'*Arabidopsis thaliana*, d'après la base de données Repbase [94].

la Mouse Transposon Insertion Database (mouse.ccgb.umn.edu/transposon/) [Roberg-Perez03]. Il existe trois bases de données spécialisées dans les ETs d'Arabidopsis : Arabidopsis Transposable Element Database [123] ; Arabidopsis thaliana Integrated Database [160] (Figure 2.2) et AtRepBase (nucleus.cshl.org/protarab/AtRepBase.htm).

AtRepBase est la première base de données qui ait relaté l'existence des hélitrons que nous étudions. Malheureusement, depuis 2002, cette base n'est plus mise à jour. Repbase est une base non spécifique qui a été créée en 1998 par Jerzy Jurka (Figure 2.2) [93, 94]. Seule Repbase a été utilisée dans ce projet, car c'est la seule base qui propose les séquences et les positions des séquences des hélitrons dans le génome d'*Arabidopsis thaliana* (Figure 2.3). Les différentes séquences hélitroniques extraites de cette base.

2.2.3 Bases de données du transcriptome d'Arabidopsis

Le transcriptome est l'ensemble des ARNm issu de l'expression des gènes d'un type de cellule dans une condition expérimentale donnée (Figure 2.4). Les données les plus nombreuses proviennent de la technologie des puces à ADN. Nous nous sommes en particulier intéressés aux données différentielles : pour un même gène, deux ARNm sont prélevés, l'ARNm de l'expérience que l'on mesure et le même ARNm exprimé dans les conditions standard de culture (Figure 2.4). Les ARNm sont réverse-transcrits en ADN avec un marquage par un fluorochrome rouge pour l'ARNm de l'échantillon testé et vert pour l'ARNm de l'échantillon normal. Ils s'hybrident avec un ADN complémentaire fixé à la paroi de la puce. L'intensité de la couleur lue au laser indique la sur-expression ou la sous-expression du gène dans la condition testée (Figure 2.3) [61].

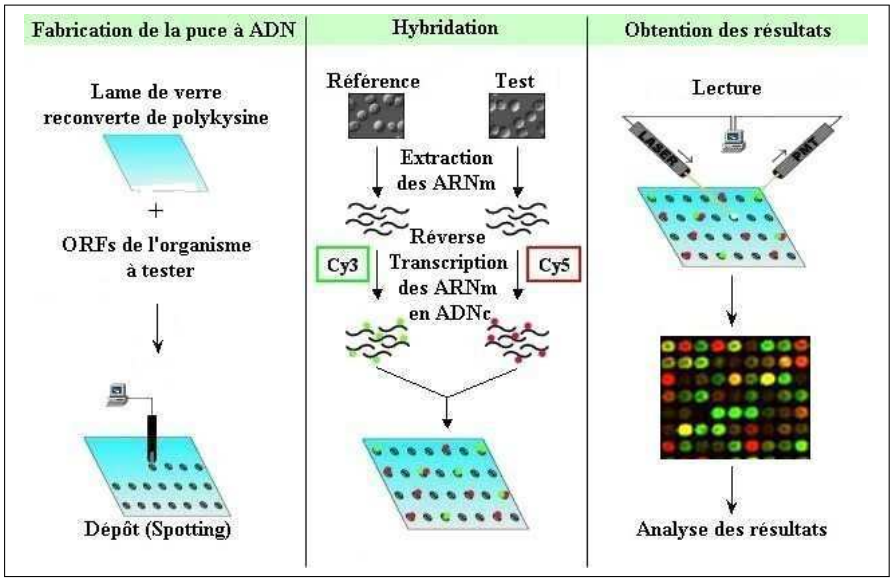


FIG. 2.4 – Technique de lecture de l'expression des gènes par une puce à ADN [61].

Il existe de nombreuses bases de données sur le transcriptome d'*Arabidopsis thaliana* (Figure 2.2). Nous avons privilégié dans nos études les données provenant de Genevestigator [212]. C'est une base qui intègre les résultats de plusieurs autres bases telles qu'AFGN, NASCArrays, Gene Expression Omnibus, FGCZ, Gruissem, Lab ArrayExpress (Figure 2.2). Elle permet de vérifier les profils d'expression dans de nombreuses conditions de

stress mais aussi dans les différents organes et les stades de développement de la plante.

Nous présentons les différentes méthodes bioinformatiques utilisées. Elles sont principalement regroupées en trois thèmes : la recherche de motifs, la classification et l'optimisation combinatoire.

2.3 Modélisation des séquences

La recherche d'éléments transposables repose dans cette thèse sur l'établissement d'un modèle de séquence qui est ensuite recherché dans l'ensemble des séquences chromosomiques d'*Arabidopsis thaliana*. Nous décrivons rapidement dans cette section les bases des modèles syntaxiques ainsi que des structures d'index qui seront utilisées pour rechercher efficacement ces modèles dans les génomes.

2.3.1 Théorie des langages formels

2.3.1.1 Définitions

Définition 1 (alphabet, mot, langage) :

- Un alphabet A est un ensemble non vide de symboles (ou lettres). Pour l'ADN, on utilise généralement : $A = \{a, t, c, g\}$ et pour les protéines l'alphabet des 20 acides aminés.
- Un mot défini sur l'alphabet A est une suite finie de symboles de A. Un mot u de longueur k ($|u| = k$) est noté $a_1a_2...a_k$. ε est le mot vide de longueur 0. $u.v$ est la concaténation de deux mots u et v de longueur n et m : $(a_1, a_2, ..., a_n).(b_1, ..., b_m) = (a_1, a_2, ..., a_n, b_1, ..., b_m)$ de longueur $n + m$. L'ensemble des mots possibles est noté $A*$.
- Un langage est un ensemble de mots. $A+$ est l'ensemble des mots de longueur supérieure ou égale à 1 ; $A^* = A \cup \varepsilon$. Nos modèles seront des langages. Il existe plusieurs moyens de représenter un langage. Nous utiliserons principalement les grammaires formelles.

Définition 2 (grammaire) :

Une grammaire formelle est un quadruplet $G = \{A, N, S, P\}$ tel que :

- A est l'alphabet des terminaux (par exemple $A = \{a, t, c, g\}$ pour l'ADN).
- N est l'alphabet des non-terminaux : les symboles utilisés pour la génération des mots.
- P est l'ensemble des règles de réécriture ou de production des mots du langage de la forme : $u_i \rightarrow u_j$ où u_i et $u_j \in N \cup A^*$.
- $S \in N$ est le symbole de départ ou l'axiome. Ce symbole est le point de départ de la génération des mots sur lequel on applique les règles de P jusqu'à obtenir une suite de terminaux.

Définition 3 (dérivation, langage généré par une grammaire) :

- Un mot v dérive d'un mot u en une étape (noté $u \Rightarrow v$) dans une grammaire G si et seulement si : il existe 2 mots x et y et une règle $u' \Rightarrow v'$ de P telle que $u = x.u'.y$ et $v = x.v'.y$. Une forme v peut être dérivée d'une forme u en plusieurs étapes où $u \Rightarrow v$, si v peut être obtenue de u par 0, 1 ou plusieurs dérivations en une étape.
- Le langage $L(G)$ généré par une grammaire $G = \{A, N, S, P\}$ est l'ensemble des mots sur A (formés uniquement de symboles terminaux) qui peuvent être dérivés à partir de S : $L(G) = \{v \in A^*/S \Rightarrow v\}$. Par exemple, la grammaire suivante définit un langage de palindromes biologiques : $G = \{A, N, S, P\}$ avec $A = \{a, t, c, g\}, N = \{S\}$ et $P = \{S \Rightarrow aSt|tSa|cSg|gSc|\varepsilon\}$. Cette grammaire peut créer le palindrome *cgacgtcg* grâce aux dérivations suivantes (les nucléotides introduits par chaque dérivation sont

colorés en rouge) : $S \Rightarrow cSg \Rightarrow cgScg \Rightarrow cgaStcg \Rightarrow cgacSgtcg \Rightarrow cgacgtcg$.

Noam Chomsky a proposé une classification des langages, la hiérarchie de Chomsky [33], basée sur leur pouvoir d'expression des structures sur les séquences.

2.3.1.2 Hiérarchie de Chomsky

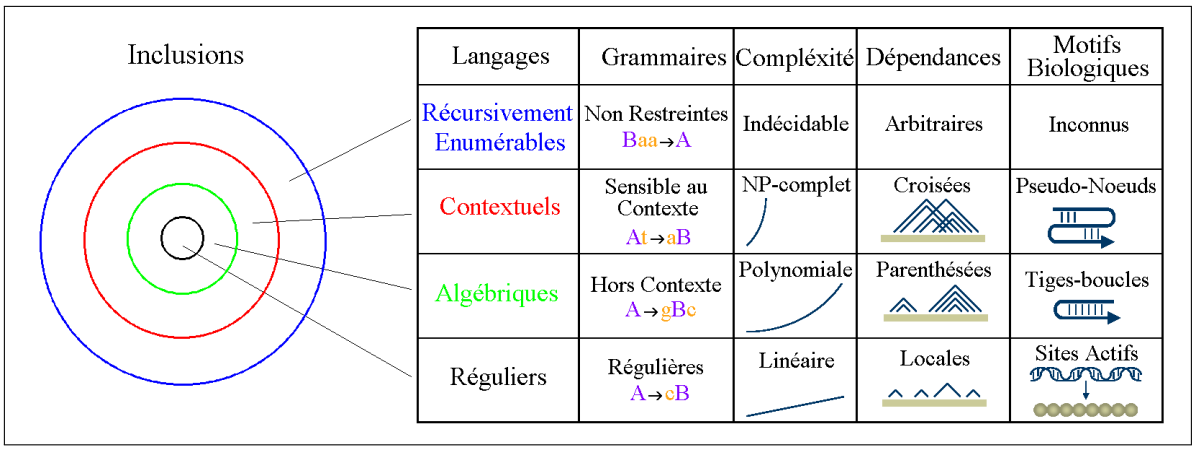


FIG. 2.5 – Hiérarchie de Chomsky et les langages sur les génomes [183]. Le tableau énumère les 4 types de langages, leur inclusion respective et les motifs biologiques reconnus par ces langages.

La hiérarchie de Chomsky [33] propose une classification des langages en 4 classes : les langages réguliers, algébriques, contextuels et récursivement énumérables. Ces langages sont emboîtés les uns dans les autres (Figure 2.5). Une classe correspond à un ensemble de propriétés qui permet de distinguer de façon très nette à la fois l'expressivité des langages qu'elle contient et la difficulté des problèmes d'analyse associés. Les langages réguliers sont définis par des grammaires où le membre gauche de chaque règle est constitué d'un seul symbole non terminal et le membre droit d'un symbole terminal, éventuellement suivi par une seule lettre non terminale. Les langages algébriques sont définis par une grammaire algébrique (ou encore hors contexte) où le membre de gauche de chaque règle est constitué d'un seul symbole non terminal (Figure 2.5). La grammaire du paragraphe précédent qui permet d'écrire des palindromes biologiques est une grammaire algébrique. Les langages contextuels sont définis par une grammaire où le contexte gauche d'une règle à une taille inférieure ou égale à son contexte droit. Les langages récursivement énumérables constituent la classe la plus générale des langages. Ils ne seront pas utilisés dans cette thèse.

2.3.2 Recherche de motifs

L'algorithmique des mots est une discipline ancienne qui connaît un développement accru avec l'émergence des nouvelles sources de séquences : séquences de génomes bien sûr, mais aussi séquences textuelles des documents disponibles sur internet. En particulier se pose le problème de la recherche d'un motif donné dans une grande séquence, problème dont la résolution passe par la mise au point de techniques d'indexation de tous les mots

présents dans cette séquence. Nous avons ainsi exploité la structure de données d'arbre des suffixes, structure bien adaptée à la recherche de répétitions dans les séquences pour la recherche systématique d'hélitrons dans le génome complet d'*Arabidopsis thaliana*. Nous avons aussi utilisé des matrices de poids pour détecter les motifs de liaison aux facteurs de transcription dans des promoteurs. Avant de décrire de ces aspects, nous commençons par introduire un minimum de vocabulaire et l'algorithme naïf de recherche de motifs dans une séquence.

- Soit un texte T de longueur n et un mot M de longueur m (généralement $n \gg m$).
- On les représente par des tableaux $M[1...m]$ et $T[1...n]$.
- Le mot M apparaît de manière exacte dans le texte T à la position j si : $1 \leq j \leq n-m$ et si $\forall i$ tel que $0 \leq i \leq m-1$ et $T[i+j] = M[i+1]$. Ce mot M est alors facteur du texte T.
- Un mot x est un facteur du mot y s'il existe deux mots u et v tel que $y = u.x.v$.
- Si $u = \varepsilon$ alors x est un préfixe de y, si $v = \varepsilon$ alors x est un suffixe de y.
- $N[1...n]$ et $M[1...m]$ sont deux mots répétés exacts si et seulement si $n = m$ et $\forall i$ tel que $1 \leq i \leq$, $N[i] = M[i]$.
- Une répétition maximale exacte est une paire d'occurrence d'un mot répété exact, de taille n dans un texte T, encadrée par deux contextes différents. Les deux occurrences sont maximales exactes si et seulement si elles sont aux positions i et j et que $T[i-1] \neq T[j-1]$ et $T[i+n+1] \neq T[j+n+1]$. Autrement dit, deux répétitions d'un mot sont maximales exactes si et seulement si leur contexte gauche et leur contexte droit sont différents (on ne peut pas rajouter une lettre à gauche ou à droite sans perdre la répétition exacte) [76].

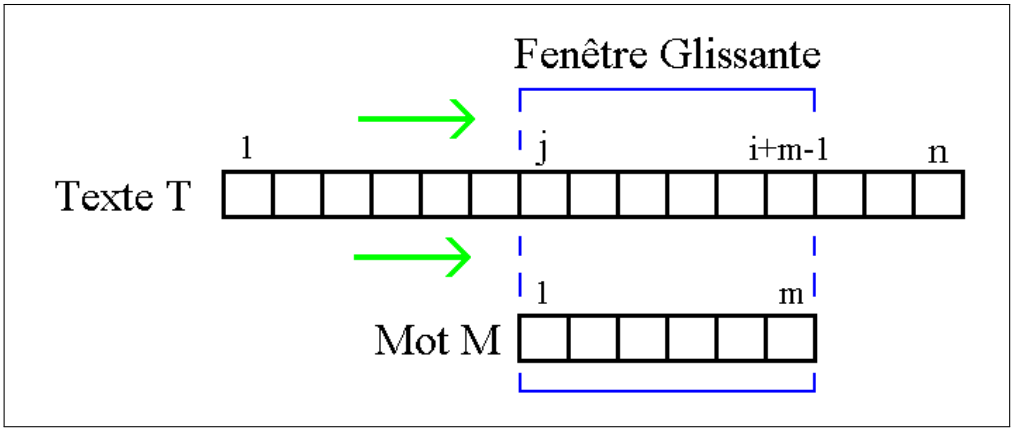


FIG. 2.6 – Représentation graphique de l'algorithme naïf. La comparaison est faite caractère par caractère. Les flèches vertes indiquent le sens de déplacement de la fenêtre glissante.

L'algorithme naïf de recherche de mots dans un texte consiste à faire glisser une fenêtre de la taille de mot M sur le texte T (Figure 2.6). Plus formellement, à chaque position j du texte on compare $T[j...(j+m-1)]$ à $M[1...m]$, la comparaison échoue si $\exists i$ $0 \leq i \leq m-1$ tel que $T[j+i] \neq M[i-1]$ (avec $0 \leq i \leq m-1$). Autrement dit, la comparaison réussit si $\forall i$ tel que $0 \leq i \leq m-1, T[j+i] = M[i+1]$. La complexité de cette algorithme est $O(n.m)$. Cet algorithme devient trop coûteux si n et m sont grands, du fait de cette dépendance multiplicative.

L'accélération de la recherche exacte ou approchée d'un mot dans le texte peut s'effectuer selon 2 approches différentes : le prétraitement du mot à rechercher et le prétraitement du texte [40, 76]. Lorsque le texte est connu, comme c'est la cas des génomes, il est préférable de prétraiter celui-ci car les algorithmes de recherche deviennent ensuite indépendants de la taille du texte.

2.3.2.1 Prétraitement du texte : Arbre des suffixes

Cette approche transforme le texte linéaire à explorer en une structure de données efficace pour la recherche de motifs [76]. Il existe plusieurs structures de données d'indexation d'un texte appropriées à la recherche de motifs tel que le DAWG (Directed Acyclic Word Graph) [41], le tableau des suffixes [135, 1], le vecteur des suffixes [168] ou l'oracle des facteurs [124]. Nous utilisons l'arbre des suffixes [207, 203, 76] une structure de données particulièrement souple pour la recherche de motifs.

L'arbre des suffixes est une structure de données décrite en 1973 par Weiner pour rechercher des répétitions de mots dans un texte (Figure 2.7) [207]. De nombreux outils de recherche de motifs et/ou de séquences répétées dans les séquences biologiques ont été créés à partir de cette structure telle que Reputer [116], MUMer [48, 117] et STAN [154]. Nous avons utilisé STAN pour la modélisation et la recherche des hélitrons.

Un arbre des suffixes est d'abord un arbre au sens mathématique du terme, c'est-à-dire un graphe acyclique orienté. Tout arbre est composé de nœuds et d'arcs reliant ces nœuds. Il existe plusieurs types de nœuds : la racine est le nœud d'origine (ou source) qui ne possède pas de parent (aucun arc n'arrive à ce nœud). Les feuilles sont des nœuds qui n'ont pas de fils (aucun arc ne sort de ces nœuds). Les nœuds internes sont les autres nœuds. Pour un texte T de taille n (n entier non nul), l'arbre des suffixes de T possède les caractéristiques suivantes [76] :

- n feuilles composent l'arbre, une pour chaque position de la séquence.
- chaque arc est étiqueté par un facteur de la séquence.
- tous les nœuds internes sont au minimum de degré 2 (au minimum 2 arcs sortants).
- tous les arcs sortant d'un même nœud commencent par une lettre différente.
- soit v un nœud, les feuilles du sous-arbre v donnent le nombre et la position des occurrences du facteur f_v allant de la racine au nœud v.
- le nœud v, ancêtre commun le plus proche de deux autres nœuds a et b, représente le plus long préfixe commun aux deux suffixes a et b.

La recherche d'un mot dans l'arbre des suffixes a une complexité proportionnelle à la taille du mot ($O(m)$) au lieu de celle du texte ($O(n)$). Par exemple, si le mot TT est recherché dans l'arbre sur ATTGAC (Figure 2.7), il suffit de lire sur l'arc sortant de la racine la lettre T, et depuis ce nouveau nœud de rechercher la lettre T dans l'un de ses arcs sortants. Une lecture directe des feuilles du sous-arbre de ce fils permet de connaître le nombre et la position du mot TT dans le texte. L'arbre des suffixes permet aussi de rechercher tous les mots répétés d'un texte, le plus long préfixe commun à deux mots du texte ou des mots répétés en tandem.

De nombreux travaux ont amélioré la construction et l'utilisation de cet arbre [143, 203, 115]. La meilleure implémentation de la construction de l'arbre des suffixes est celle de Kurtz [115]. Cette implémentation est une optimisation de l'algorithme de Ukkonen qui permet de construire l'arbre en temps $O(n)$ et *on line* [203].

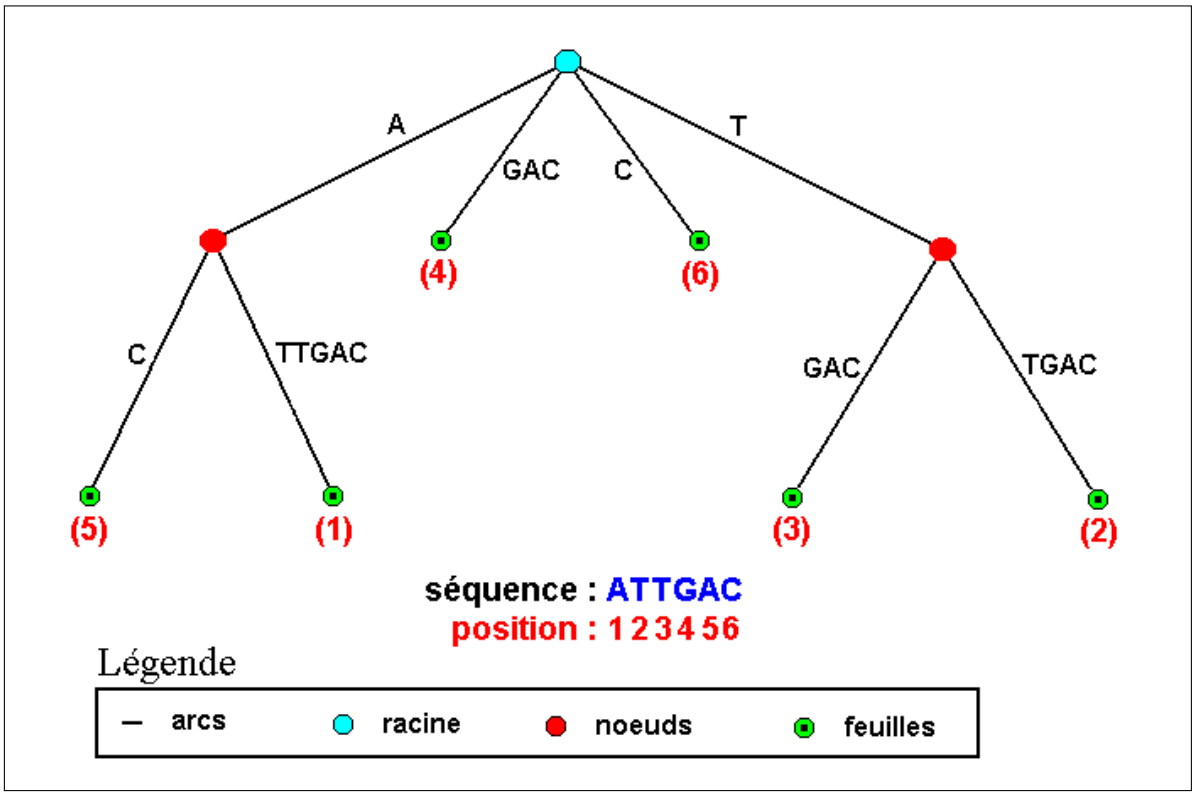


FIG. 2.7 – Arbre compact des suffixes de la séquence ATTGAC. Les arcs sont étiquetés par les facteurs de la séquence. A chaque feuille est indiquée la position du suffixe dans la séquence correspondant au mot lu depuis la racine jusqu'à la feuille.

2.3.2.2 Grammaires SVG

Pour rechercher un motif biologique complexe, tels que les hélitrons, dont le modèle de la littérature mélange structure primaire et secondaire (Figure 1.19), les langages réguliers ne suffisent pas. Néanmoins, les langages contextuels, qui sont capables de détecter n'importe quel motif biologique connu, ne peuvent être utilisés dans toute leur généralité en bioinformatique à cause de leur complexité algorithmique (Figure 2.5). Les grammaires à variable de chaînes ou SVG (String Variable Grammar) ont été mis au point à partir des DCG (Definite Clause Grammars) et des grammaires indexées [182, 184], pour définir une sous-classe intermédiaire entre les langages contextuels et non contextuels. Les grammaires SVG sont utilisées dans de nombreux travaux de modélisation de motifs ADN et de régulation génique [181, 80, 130, 34].

Les grammaires SVG introduisent le concept de variable de chaîne qui peut être associée à une chaîne de caractères quelconque dans une recherche de motifs. Une variable de chaîne est représentée par un identifiant tel que la lettre X. Ces variables peuvent être contraintes en taille en utilisant la syntaxe $X{:}[a,b]$, où a et b sont les limites inférieures et supérieures de la taille de X. Un token (une chaîne de caractères ou un caractère) ou une variable de chaîne peut être préfixée par l'opérateur $\sim$ qui précise à l'analyseur de rechercher le complément inverse. Ainsi, le motif $X{\sim}X$ permet de rechercher des palindromes. De plus, les erreurs de substitution, définies par le symbole ":" suivi du nombre d'erreurs

maximales permises, sont autorisées dans la définition d'une variable. Par exemple, $X \sim X{:}a$ permet de rechercher des palindromes avec a erreurs de substitution.

Les grammaires SVG ont donné lieu à l'implémentation de deux analyseurs appelés GenLang [184] et STAN [154]. Néanmoins, le code GenLang n'est plus maintenu et sa complexité le limite à la recherche de séquences de quelques mégabases. Au début de cette thèse, nous avons utilisé Genlang [184]. Certaines recherches telles que la recherche de palindromes dans les hélitrons ou les hélitrons dans le génome entier d'*Arabidopsis thaliana* prenaient alors plusieurs heures, voire plusieurs jours. A cause de cette lenteur, nous avons abandonné cet analyseur. Nous avons utilisé le programme STAN, car il permet d'avoir la souplesse des langages avec variables de chaîne, similaire à Genlang, grâce à un analyseur Prolog et la rapidité d'exécution de l'arbre des suffixes (www.irisa.fr/symbiose/STAN) [154].

2.3.3 Application de STAN à la modélisation des séquences

Nous avons utilisé l'analyseur STAN et les grammaires SVG pour la détection de la plupart des motifs biologiques de cette thèse : la recherche des éléments transposables ou de parties d'ETs dans un génome entier et la recherche de motifs de transcription dans des promoteurs.

2.3.3.1 Recherche exhaustive des éléments transposables dans un génome entier

La recherche des éléments transposables ou des motifs caractéristiques des éléments transposables a été exclusivement réalisée avec STAN [154]. En plus de la recherche des hélitrons, nous l'avons utilisé pour les MITEs Emigrant [69]. Par exemple, la recherche de cet élément s'écrit : $X{:}[2]{-}TIR5'{-}x(0,1000){-}TIR3'{-}X$. Dans cette grammaire TIR5' et TIR3' sont remplacés par les séquences des TIRs 5' et 3'. Les deux X représentent les TSD qui sont les deux séquences répétées entourant les TIRs. La recherche de ce motif dans le génome d'*Arabidopsis thaliana* a fourni toutes les occurrences de ce pattern en moins de 3 minutes sur le cluster de PC de la Ouest-génopole (www.irisa.fr/symbiose/STAN). Nous avons aussi employé STAN pour la recherche exhaustive de structures secondaires (Exemple : tige-boucle subterminale des hélitrons) dans la séquence interne des transposons. Cette recherche par modélisation des hélitrons est l'un des principaux résultats de cette thèse, elle sera abordée dans le chapitre suivant "Résultats et Discussion".

2.3.3.2 Recherche des motifs de liaison aux facteurs de transcription

Des motifs de régulation permettent le contrôle de l'expression d'un gène dans l'espace et dans le temps : un gène est exprimé à certains moments de la vie de la cellule, et dans certains compartiments de la cellule. Ces motifs sont essentiellement présents dans le promoteur (Figure 2.8) en amont du gène [36, 139, 193], mais certains motifs ont été découverts dans les régions transcrites non traduites de l'ARNm [55].

La base de la découverte de motifs de régulation *in silico* est la recherche de mots exceptionnels dans le promoteur : mot avec une fréquence d'apparition élevée à l'intérieur des zones promotrices de gènes ayant des profils d'expression similaires, mais avec une fréquence d'apparition faible dans l'ensemble des promoteurs des gènes pris aléatoirement [175].

Dans ce contexte de recherche de mots répétés, les éléments transposables sont généralement enlevés du promoteur étudié, car leur fréquence d'apparition en dehors et dans les promoteurs est élevée.

Deux grandes méthodes permettent de trouver des motifs de régulation dans les promoteurs : la détection *ab initio* de ces motifs ou la recherche de ces motifs à partir de banques de données de motifs. La détection de motifs est principalement basée sur deux principes : la probabilité d'apparition d'un mot, et l'empreinte phylogénétique [175].

L'approche basée sur les probabilités d'apparition est une méthode *ab initio*. On calcule la fréquence de tous les mots présents dans la séquence et la probabilité d'obtenir une telle fréquence pour une séquence aléatoire [67]. Les deux programmes les plus connus qui utilisent cette méthode sont MEME [11] et MotifSampler [198].

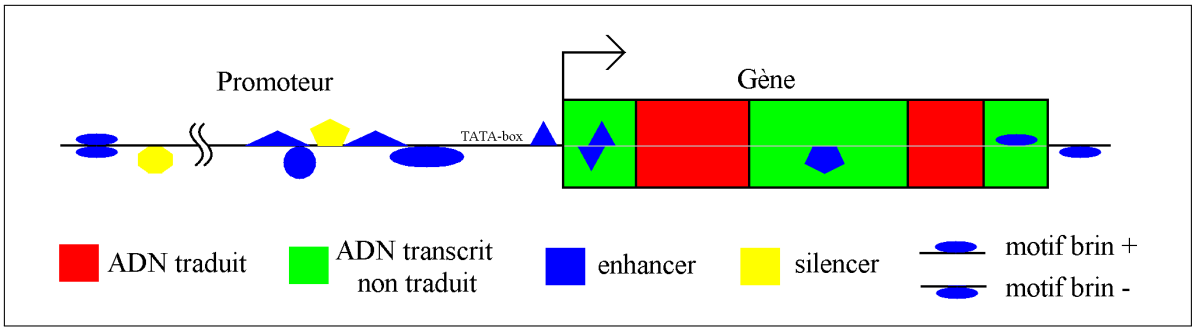


FIG. 2.8 – Schéma récapitulatif des emplacements possibles des motifs de liaisons des facteurs de transcription pour un gène eucaryote [36, 193]. Un enhancer (silencer) est un motif de liaison aux facteurs de transcription qui augmente (diminue) la transcription.

La méthode basée sur l'empreinte phylogénétique aligne la séquence promotrice donnée avec une séquence orthologue dans un autre génome [24, 175]. Cette méthode suppose que les motifs de régulation sont conservés au cours de l'évolution des espèces proches. Elle est de plus en plus utilisée avec la disponibilité croissante de séquences de génomes.

La méthode basée sur les banques de motifs utilise des motifs de régulation connus expérimentalement. Ces motifs sont recherchés dans le promoteur sous deux formes : le mot lui-même ou une matrice de poids [175]. Les bases de données de motifs de régulation tel que TRANSFAC [209] ou PlantCARE (Figure 2.2) [129] contiennent les motifs sous la forme de matrice de positions.

L'interface TOUCAN (homes.esat.kuleuven.be/ saerts/software/toucan.php) [2] permet d'utiliser les différentes méthodes telles que MotifSampler [198]. Ce programme et plusieurs bases de données de motifs de régulation telles que AGRIS (Figure 2.2) [45] ou TRANSFAC (www.gene-regulation.com/pub/databases.html) [140] ont été utilisés lors de ce projet pour réaliser des analyses croisées des différents promoteurs à notre disposition et confronter leurs différentes prédictions.

2.4 Classification des séquences

Au cours de ma thèse, j'ai cherché à classer les différentes familles d'hélitrons ou les différentes copies d'une même famille, en fonction de leurs liens évolutifs, mais aussi en fonction de leurs caractéristiques morphologiques telles que la séquence de l'hairpin ou l'absence/présence d'un ORF, etc. La variabilité de séquence des hélitrons a rendu l'utilisation de méthodes phylogénétiques classiques très difficiles à utiliser [98]. Lorsque cela était possible j'ai utilisé la méthode du Neighbour-Joining [177] précédée d'un alignement multiple réalisé avec ClustalW [200].

Pour classifier les séquences des hélitrons, j'ai utilisé un logiciel de classification ascendante hiérarchique appelé CHAVL (Classification Hiérarchique par Analyse de la Vraisemblance des Liens) [128]. Cette méthode de classification utilise un tableau de données croisant l'ensemble des objets O (ou individus) et l'ensemble des attributs A (ou des variables descriptives). CHAVL peut considérer plusieurs types de variables : attributs booléens, variables qualitatives ordinales, nominales, numériques et préordonnances. Au cours de ce doctorat, nous avons essentiellement utilisé CHAVL sur des données numériques : nous avons codé dans les variables le nombre de fois où une caractéristique biologique donnée était détectée dans un hélitron. La méthode CHAVL permet de classifier aussi bien les individus que les variables. CHAVL est basé sur la notion de la vraisemblance d'une ressemblance entre les objets à classifier.

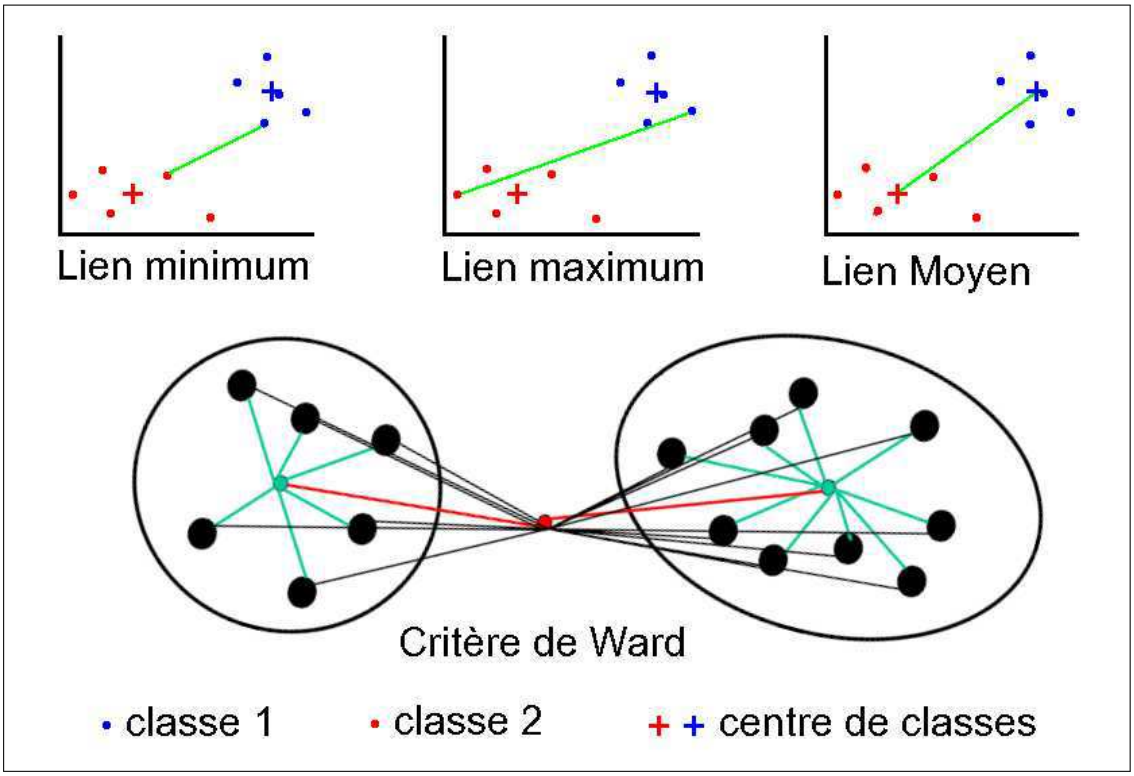


FIG. 2.9 – Différents critères d'agrégation des classes : lien minimum, maximum, moyen et critère de Ward. Les traits verts clairs pour les liens minimum, maximum et moyen représentent la distance entre les classes 1 et 2 pour ces 3 liens. Les traits verts du critère de Ward représentent l'inertie intra-classe, les traits rouges représentent l'inertie inter-classe.

La mesure de similarité s'exprime dans une échelle probabiliste et tient compte de la représentation mathématique, statistique et logique des objets. La vraisemblance est évaluée sous une hypothèse d'absence de liens (H_0) entre les classes définissant les objets. L'indice brut de similarité entre deux objets α et β est défini par : $S_{\alpha\beta} = Card(O_\alpha \cap O_\beta)$ où O_α et O_β dépend du type de variables. Nous avons utilisé les variables booléennes et les variables numériques. Les deux types de variable utilisent le même indice de similarité si les valeurs booléennes sont convertis en 0 et 1 (respectivement FAUX et VRAI).

Le critère d'agrégation détermine la mesure entre les différentes classes d'objets. Quatre critères sont principalement utilisés : le lien minimum (simple linkage), le lien maximum (complete linkage), le lien moyen (average linkage ou UPGMA) et le critère de Ward. Le lien minimum (maximum) est la distance minimale (maximale) qui sépare deux objets de deux classes. La Figure 2.9 montre les distances entre deux classes pour les trois premiers liens et montre les distances intra-classes et inter-classes pour la critère de Ward. Ce critère minimise la première distance et maximise la seconde.

L'algorithme de classification est l'algorithme simple de classification ascendante hiérarchique (CAH). A l'initialisation de l'algorithme, chaque objet est dans une classe différente et un indice de similarité S est calculé entre chaque paire de classes. L'algorithme de CAH va agréger à chaque pas les classes dont les indices S sont les plus proches, selon le critère d'agrégation, et va créer une nouvelle classe qui est la fusion des deux précédentes (Figure 2.10). L'algorithme calcule ensuite l'indice entre la nouvelle classe et les autres classes existantes. Au fur et à mesure de l'avancement de l'algorithme, un arbre contenant l'historique de la classification est créé et permet de retrouver au plus $n - 1$ partitions finales (pour n objets) (Figure 2.10).

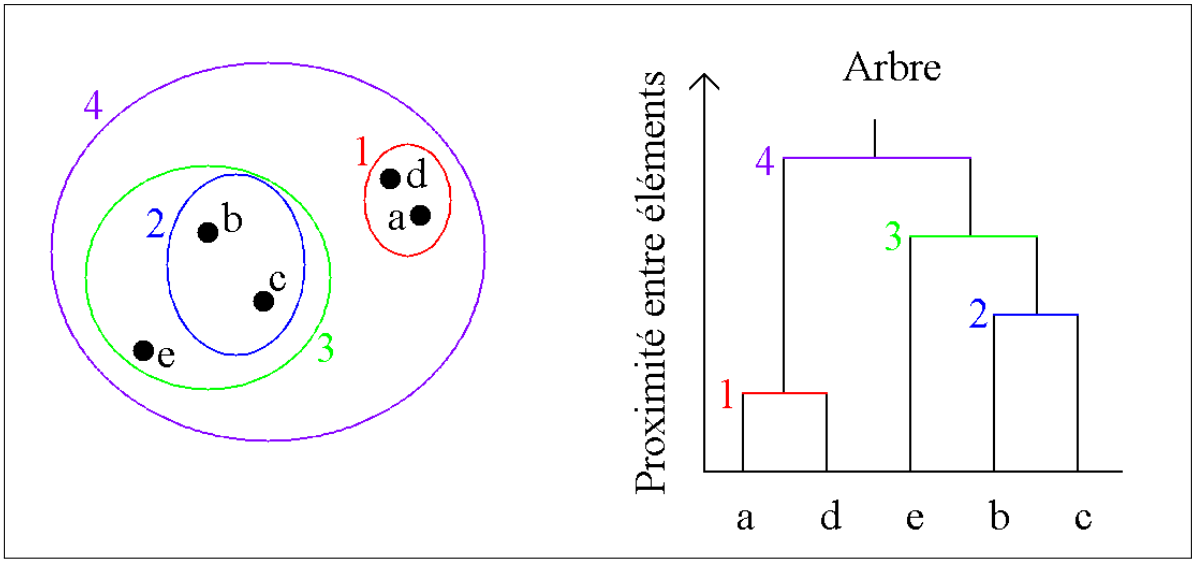


FIG. 2.10 – Exemple de classification ascendante hiérarchique. Les objets sont représentés par les lettres a, b, c, d et e. Ces objets ont leurs coordonnées dans le plan xy comme variable, et le critère d'agrégation des objets est simplement la distance séparant les objets sur ce plan. L'arbre représente l'historique d'agrégation des objets en partant des premières agrégations en bas de l'arbre pour finir par la dernière agrégation en haut de celui-ci.

Pour aider à l'interprétation de la classification, CHAVL fournit 3 indices :

- La statistique 'globale' est une mesure de cohésion entre une partition obtenue à un niveau et la similarité entre les couples d'objets (cette similarité étant représentée par la préordonnance définie sur l'ensemble des couples de paires d'objets). Donc cette statistique mesure en quelque sorte la qualité de la partition (plus exactement, l'adéquation entre une partition et la mesure de similarité).
- la statistique locale (à chaque niveau) permet de repérer les "nœuds significatifs". La statistique locale mesure la variation de la statistique globale à un niveau donné par rapport au niveau précédent.
- un indice de neutralité permet de distinguer les objets non classifiables vis-à-vis de la classification des autres objets.

2.5 Optimisation combinatoire de données

Un problème d'optimisation combinatoire consiste à trouver la meilleure solution dans un ensemble fini O, dit ensemble des solutions réalisables. On dispose d'une fonction f qui calcule, pour chaque solution réalisable x, un coût $f(x)$. Le but de l'optimisation combinatoire est de retrouver la solution réalisable x de coût minimal.

Trouver une solution optimale dans un ensemble discret et fini semble un problème facile en théorie : il suffit d'énumérer toutes les solutions, et de comparer leurs coûts pour voir quelle est la meilleure. Cependant, en pratique, le problème se confronte à deux difficultés. La première est que l'ensemble O est très grand et que l'énumération de ses éléments peut prendre un temps exponentiel en fonction de la taille du problème. La deuxième est qu'il n'existe pas toujours en pratique un algorithme simple pour énumérer les solutions de O. Dans la plupart des cas, le problème est NP-complet (Non Polynômial). En conséquence, beaucoup de méthodes ont été développées pour approcher ce type de problèmes. L'optimisation combinatoire se trouve ainsi au carrefour de plusieurs disciplines telles que :

- la combinatoire.
- l'algèbre linéaire.
- la programmation linéaire.
- la programmation en nombres entiers.
- la théorie des graphes.
- les polyèdres combinatoires.
- la complexité des algorithmes.

Au cours de cette thèse, nous avons été confronté à deux problèmes d'optimisation différents qu'il était impossible de résoudre manuellement :

- Parmi tous les couples possibles d'extrémités hélitroniques qui représentaient toutes les occurrences des hélitrons, nous avons dû choisir un jeu de couples d'extrémités qui utilise chacune des extrémités, une seule fois, et qui maximise le nombre d'occurrences pour l'ensemble des couples choisis : ce problème correspond à un problème d'affectation de tâches en optimisation combinatoire.
- A partir de toutes les séquences hélitroniques d'*Arabidopsis thaliana*, nous avions un très grand ensemble de couples possibles pour représenter ces séquences, nous avons voulu déterminer le jeu de couples minimum qui représente toutes les séquences : ce problème est un problème de couverture (utilisation de la théorie des graphes et de la programmation linéaire).

2.5.1 Algorithme de Munkres

Le problème d'affectation consiste à assigner différentes tâches à des individus afin de réduire au minimum le coût total des tâches. Chaque individu peut exécuter chacune des tâches, mais chacune à un coût différent. S'il y a n tâches et n individus, on utilise une matrice C de coût $(n \times n)$, où chaque élément (i, j) représente le coût d'affecter la $i^{ième}$ tâche au $j^{ième}$ individu. Chaque individu peut exécuter seulement une tâche et chaque tâche est assignée à seulement un individu.

Par défaut, l'algorithme exécute la minimisation des coûts pour l'ensemble des couples choisis dans les éléments de la matrice.

L'algorithme Hongrois (ou de Munkres) est un exemple d'algorithme d'optimisation combinatoire qui résout les problèmes d'affectation des tâches en temps $O(n^3)$. Cet algorithme a été publié par Harold Kuhn en 1955 [113] et amélioré en 1957 par James Munkres [152]. L'algorithme d'assignement de tâches de Munkres peut être étendu aux tableaux rectangulaires [18].

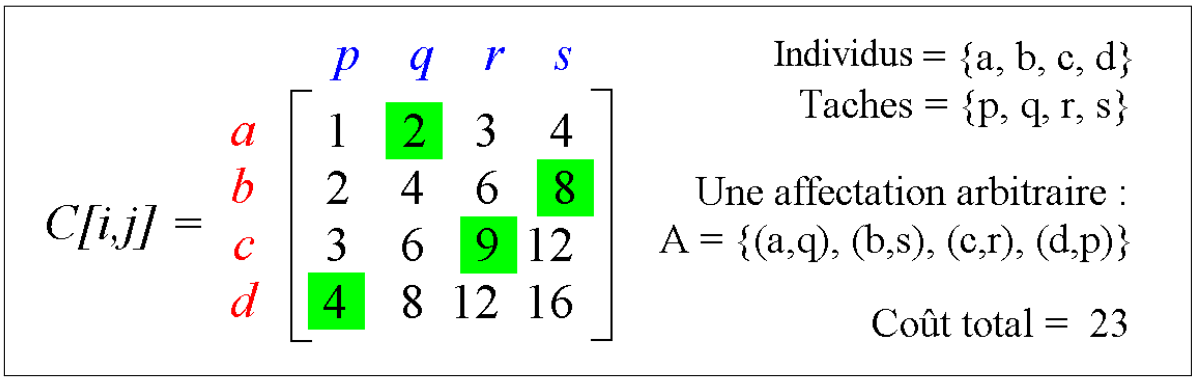


FIG. 2.11 – Exemple d'affectation arbitraire. L'individu a est assigné à la tâche q, l'individu b est assigné au travail s etc. Le coût total de cette affectation est 23.

2.5.2 Problème de couverture minimale

Le problème de couverture (Set Cover) est un problème classique en informatique et en théorie de la complexité. Un problème de couverture consiste à trouver un jeu minimum de sous-ensembles qui contient tous les éléments présents dans la totalité des sous-ensembles. Additionnellement, certains problèmes rajoutent des coûts à chaque sous-ensemble, il faut alors minimiser le coût du jeu minimum des sous-ensembles. Ce problème est l'un des 21 problèmes de Karp prouvé être NP-complet en 1972 [38]. Plus formellement, un problème de couverture minimale est décrit comme :

- l'ensemble des éléments (ou Univers) $O = \{o_1, o_2, ..., o_n\}$.
- l'ensemble des sous-ensembles $S_1, S_2, ..., S_k \subseteq O$.
- et l'ensemble des coûts $c_1, c_2, ..., c_k$ (pour un problème de couverture sans coûts, $c_i = 1, \forall i \in I$).
- on recherchons le jeu $I \subseteq \{1, 2, ..., k\}$ qui minimise $\sum_{i \in I} c_i$ avec $\cup S_i = O$.

De nombreux problèmes biologiques ont été formalisés en problème de couverture dont les interactions protéines-protéines [81], la sélection minimale d'amorces pour reconnaître un ensemble de microorganismes [10] et le positionnement des amorces LR-PCR sur le génome de *Staphylococcus aureus* [5].

L'algorithme glouton de couverture minimale donne une solution avec une complexité algorithmique polynômiale. L'algorithme choisit à chaque étape le sous-ensemble qui couvre le maximum d'éléments de O. L'algorithme est présenté ci-dessous :

L'algorithme glouton peut trouver la solution optimale, mais dans de nombreux cas, ce n'est absolument pas garanti. Par exemple la Figure 2.12 montre un exemple de problème de couverture où l'algorithme glouton ne détecte pas la solution optimale. Dans ce cas, l'algorithme choisit le premier sous-ensemble qui contient le plus d'objets : le sous-ensemble

Algorithm 1 Algorithme glouton de Couverture minimale

Require: Univers $O = \{o_1, o_2, ..., o_n\}$
Require: Sous-ensemble $S = \varnothing$
  while O contient au moins 1 élément **do**
    Choisir le sous-ensemble S_i qui couvre le maximum d'éléments de O
    $S = S + S_i$
    $O = O - S_i$
  end while
  return le jeu de sous-ensembles S

noir qui compte 8 objets, on enlève les objets de ce sous-ensemble, l'algorithme choisit ensuite le sous-ensemble vert puis le jaune. La solution donnée par l'algorithme glouton contient trois sous-ensembles, alors que la solution optimale est de deux sous-ensembles (bleu et rouge) (Figure 2.12).

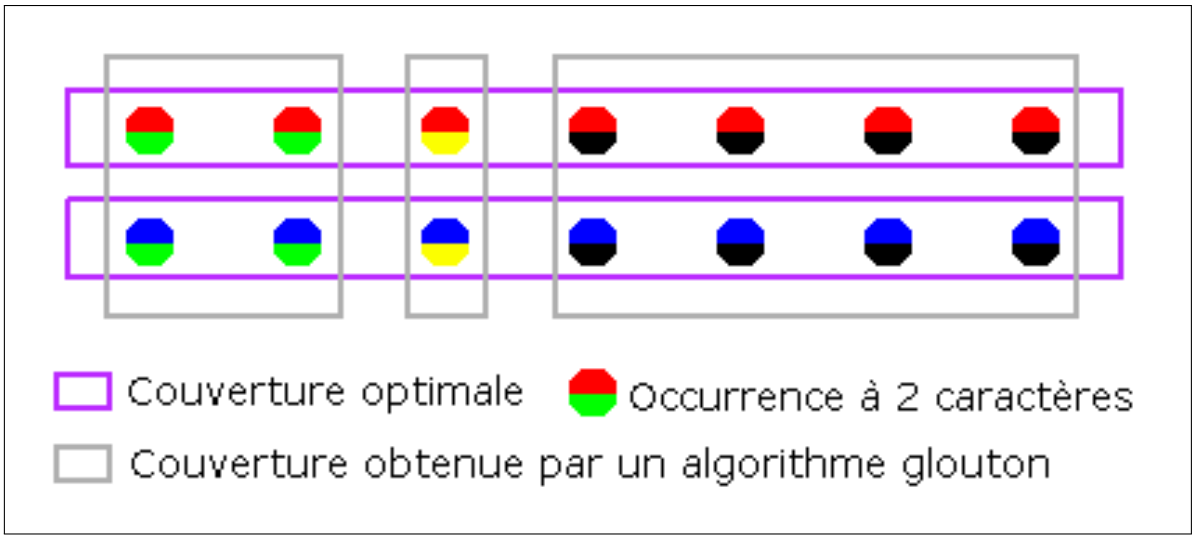


FIG. 2.12 – Exemple de solutions pour un problème de couverture minimale sans coût associé. Chaque rectangle est un sous-ensemble possible. Les rectangles rouges montrent la solution trouvée par l'algorithme glouton, les rectangles mauves montrent la solution optimale de ce problème.

Résultats et Discussions

3.1 Découverte d'un élément hélitron dans le gène Arginine DéCarboxylase 1 (ADC1)

Dans cette première section, nous introduisons l'étude à l'origine de la thèse : la découverte d'un élément transposable dans le gène ADC1 [62] et son identification comme un élément de type hélitron.

3.1.1 Analyse comparative des promoteurs des deux gènes de l'ADC

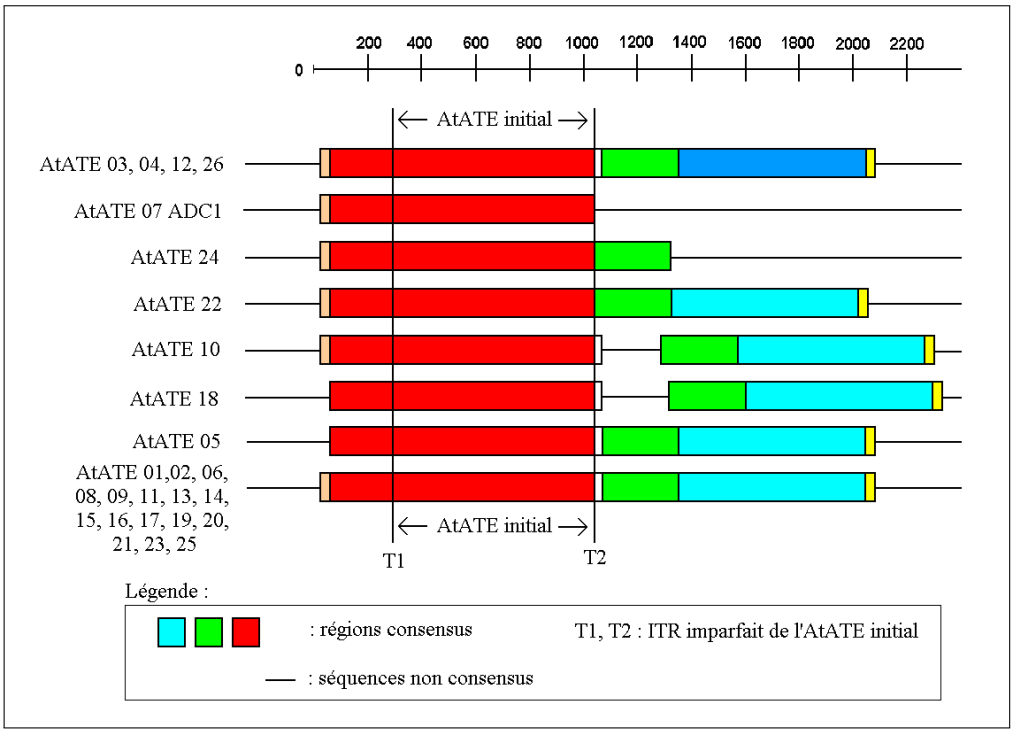


FIG. 3.1 – Structure et extension des 26 AtATEs initialement découverts [62]. L'élément transposable du promoteur ADC1 est l'AtATE 07 noté AtATE 07 ADC1. Les couleurs montrent les similarités de séquences internes entre différentes régions.

Les polyamines sont largement impliquées dans le développement et les réponses aux stress chez l'ensemble des organismes. L'ADC (Arginine DéCarboxylase) constitue une

protéine clé dans la voie de biosynthèse des polyamines. Dans le génome de la plante modèle *Arabidopsis thaliana*, l'ADC est codé par deux gènes, le premier est porté par le chromosome II (ADC1) et le second par le chromosome IV (ADC2). Bien que la protéine codée par ces deux gènes présente une homologie de plus de 80 %, l'utilisation de l'approche *promoteur : :gène rapporteur* a permis de montrer que leur profil d'expression est largement différent. L'ADC2 présente une expression plutôt généraliste, alors que l'ADC1 présente une expression plutôt de type tissu et stress spécifique. Ces observations nous ont amené à faire une analyse fine des séquences nucléotidiques des régions promotrices, pour identifier éventuellement des boîtes transcriptionnelles ou des structures originales responsables de ces différences.

L'analyse des régions promotrices des deux gènes paralogues de l'ADC chez Arabidopsis, par BLAST [3], a montré que les deux régions présentaient un faible degré de similitude et une large variété de boîtes de régulations transcriptionnelles putatives. Curieusement, la séquence promotrice de l'ADC1 contient une séquence, non seulement très répétée dans le génome, mais fortement conservée. Les analyses de cette structure ont permis de montrer qu'il s'agissait d'un élément transposable. Il est hautement répété : 26 copies complètes et 1671 copies partielles [62]. Cet élément répété, appelé initialement AtATE (*Arabidopsis thaliana* ADC1 Transposable Element) est constitué de 742 pb et présente les caractéristiques des MITEs avec des ITRs imparfaits, mais ne possède aucun TSD [62].

3.1.2 Analyse systématique des régions flanquantes d'AtATE

L'élément AtATE est un élément non-autonome, il ne possède pas d'ORFs. Nous avons alors recherché la famille d'AtATE pour découvrir éventuellement l'élément autonome à l'origine de sa création.

Nous avons procédé tout d'abord à l'analyse systématique des séquences flanquantes d'AtATE, afin de rechercher éventuellement des TSD caractéristiques des MITEs. Cette analyse a montré que les 26 éléments répétés étaient beaucoup plus grands et pouvaient s'étendre sur environ 300 pb en 5' et 1000 pb en 3', sauf pour l'élément présent dans le promoteur ADC1 (Figure 3.1). Ces nouvelles séquences flanquantes ne présentent aucune structure répétée et/ou palindromique d'un élément transposable classique. Une recherche bibliographique et un alignement de séquence avec ClustalW [200] ont alors montré une forte similarité entre l'élément AtATE étendu et l'élément non-autonome AtREP3 du type Hélitron [98].

Résumé

L'élément transposable présent dans le promoteur du gène ADC1 est un élément de type Hélitron. Il fait partie de la famille non-autonome AtREP3.

3.2 Conception d'un modèle syntaxique d'hélitron à partir de la famille AtREP3

Les précédentes études décrites sur les hélitrons chez *Arabidopsis thaliana* ont montré l'existence, en plus d'AtREP3, de cinq familles autonomes et de 29 familles non-autonomes [98, 94]. Le motif consensus connu des hélitrons est : TC en 5' bordé par le nucléotide A, et CTAG bordé par le nucléotide T avec une tige-boucle subterminale en 3' (Figure 1.19). Ce motif est un modèle trop peu sélectif pour permettre une reconnaissance précise des familles. La première étape de cette thèse a été de créer un nouveau modèle syntaxique des hélitrons et de mieux comprendre les relations entre hélitrons autonomes et non-autonomes.

3.2.1 Amélioration du modèle de recherche des AtREP3s

Généralement, les extrémités des éléments transposables sont utilisées comme signaux de reconnaissance par les protéines de transposition [39]. De fait, le mode de transposition présumé des hélitrons indique qu'ils utilisent leurs extrémités [70]. D'autre part, la séquence des AtREP3s, et des hélitrons en général, montre une grande variation des séquences internes (Figure 3.1) [98]. Par conséquent, cette séquence interne ne peut pas servir de motif consensus pour réaliser un modèle de recherche systématique des hélitrons.

Pour créer un modèle hélitronique uniquement en se basant sur les deux extrémités, il était important de déterminer d'abord la taille optimale qui permettait de discriminer les extrémités de l'ensemble des AtREP3s décrits dans le génome d'*Arabidopsis thaliana*. La séquence consensus des deux extrémités hélitroniques d'AtREP3 a été extraite de la base de données Repbase (www.girinst.org/repbase/index.html) [94].

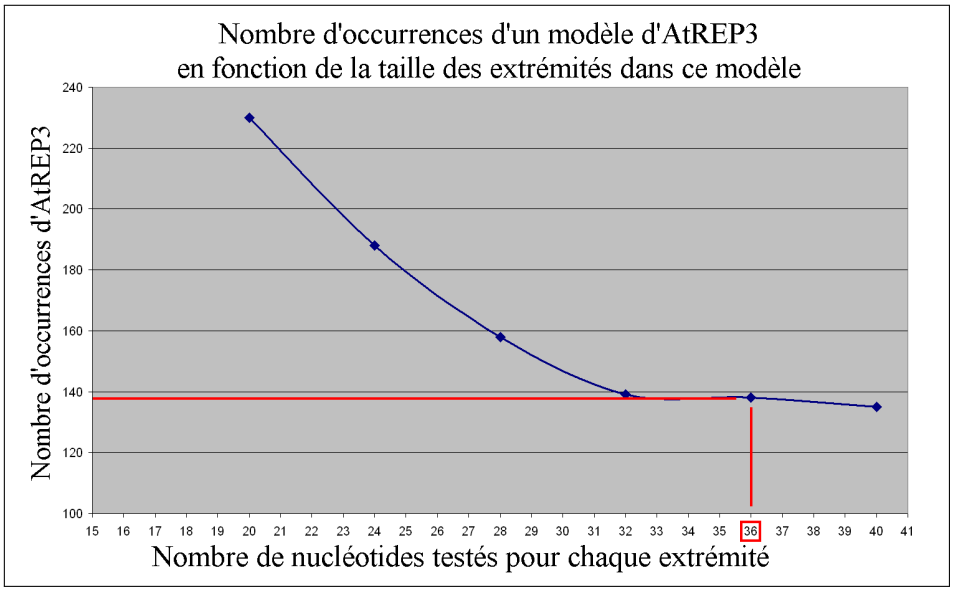


FIG. 3.2 – Nombre d'occurrences d'AtREP3 dans le génome d'*Arabidopsis thaliana* en fonction de la taille des deux extrémités. La valeur encadrée représente le nombre de nucléotides choisi. La valeur d'occurrences correspondante est similaire à celle trouvée par Kapitonov et Jurka [98].

Nous avons effectué le test de l'ensemble des modèles basés sur les séquences des extré-

mités, pour des tailles variant de 20 à 40 pb. Pour chaque test réalisé, les deux extrémités avaient la même taille. De plus pour chaque test de taille, un pourcentage d'erreur de 25 % était toléré sur les séquences, car l'analyse préliminaire des 26 AtREP3s montrait de légères substitutions ne dépassant pas le seuil de 25 % d'erreurs de substitutions sur l'ensemble de la séquence des extrémités.

La taille de la séquence interne entre les deux extrémités des 26 AtREP3s varie entre 1000 et 2100 pb [98]. Le nouveau modèle de l'hélitron AtREP3 doit autorisé une taille entre 100 et 3000 pb pour ne perdre aucune occurrence d'AtREP3. Cette structure, séquence variable bordée en 5' et 3' de séquences connues, s'écrit avec STAN de la manière suivante :

```
Extrémité_5'_25%_d'erreurs-x(100,3000)-Extrémité_3'_25%_d'erreurs
```

Enfin, chaque valeur de fréquence d'AtREP3 a été comparée à la valeur obtenue par Kapitonov et Jurka [98] et les séquences ont été alignées pour repérer les séquences détectées qui ne sont pas des AtREP3s (faux positifs).

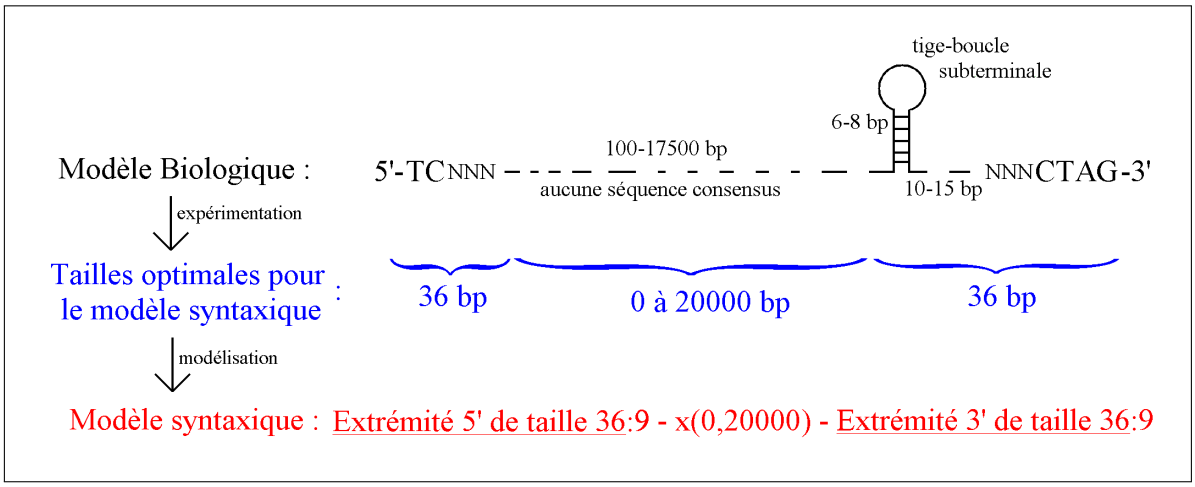


FIG. 3.3 – Correspondance entre les connaissances biologiques et le modèle syntaxique. La partie en noir représente les données de la littérature. La partie en bleu représente les informations que nous avons obtenues à partir d'expérimentations sur la taille des extrémités. Le modèle en rouge représente le modèle final de l'hélitron avec 2 extrémités de 36 paires de bases séparées par un gap variable. La valeur 9 représente le nombre maximal d'erreurs de substitution accepté par rapport au motif consensus extrait de Repbase.

Nous avons pu montrer que la taille des 36 nucléotides pour chaque extrémité était la taille optimale qui permettait d'avoir un nombre d'occurrences d'hélitrons similaire à celui publié par Kapitonov et Jurka (139 non chevauchantes comparé à 150) (Figure 3.2). Nous avons aligné les 139 séquences obtenues par le modèle d'AtREP3 et la séquence consensus d'AtREP3 de Repbase [94]. L'alignement multiple a montré que toutes les séquences étaient bien des hélitrons AtREP3 (aucun faux positif). La majorité des occurrences d'AtREP3 obtenues par STAN avait une séquence similaire à la séquence consensus et les autres occurrences montraient une large délétion de la séquence 5' subterminale, mais elles avaient toute la partie 3' similaire à l'AtREP3 consensus.

L'alignement des 139 AtREP3s a montré plusieurs insertions correspondant à des éléments transposables. Parmi ces insertions deux séquences insérées possédaient les caractéristiques des hélitrons [98]. Aucun ORF n'a été détecté dans ces séquences et leurs couples d'extrémités ne correspondaient à aucune famille d'hélitron connue. Nous avons nommé ces séquences AtREP20 et AtREP21. Après une recherche des différentes copies, les consensus des deux familles ont été déposées dans Repbase [196].

Nous avons ensuite étendu l'étude en appliquant ce modèle à toutes les familles connues d'hélitrons prélevées dans Repbase [98, 94] à l'exception des AtREP16, 17, 18 et 19 qui ne possèdent pas les caractéristiques terminales de séquences primaires des hélitrons (Figure 1.19). La taille des familles d'hélitrons varie entre 500 et 18000 bp. Le modèle définitif que nous avons retenu pour le type hélitron chez *Arabidopsis thaliana* est de la forme (Figure 3.3) :

```
Extrémité_5'_de_36pb:9-x(0,20000)-Extrémité_3'_de_36pb:9
```

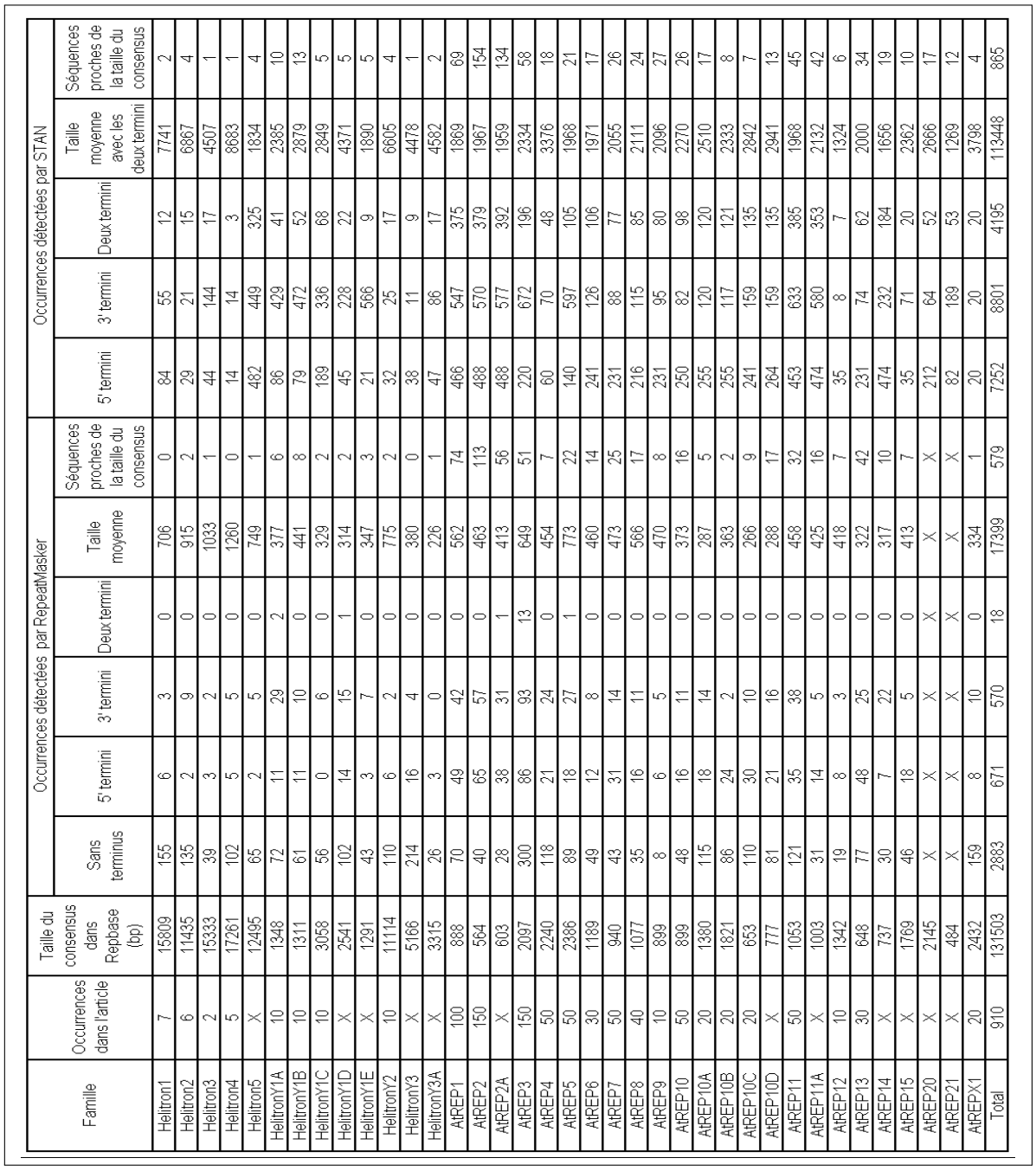
3.2.2 Comparaison de la méthode syntaxique développée avec la méthode traditionnelle de détection des éléments transposables

Nous avons voulu comparer notre méthode syntaxique avec la méthode de détection CENSOR [95, 107] utilisée par Kapitonov et Jurka [98]. Cette méthode nécessite l'utilisation de paramètres spécifiques pour être efficace et ceux-ci n'ont pas été donnés dans leur article [98]. Nous avons donc comparé les occurrences obtenues par ce modèle et les occurrences obtenues avec RepeatMasker (Figure 3.3) qui est la méthode la plus connue et une méthode proche de la méthode utilisée par Kapitonov et Jurka [98].

RepeatMasker utilise WU-BLAST ou des logiciels similaires qui réalisent un alignement de séquences ou de génomes avec une librairie d'éléments transposables connus. Nous avons utilisé la version open-3.1.6 de RepeatMasker avec la version 2.0 de WU-BLAST ainsi que la version 20061006 de la librairie d'éléments transposables qui a été téléchargée sur le site de Repbase (www.girinst.org/repbase/index.html).

Nous avons choisi de comparer la détection des hélitrons avec RepeatMasker, et la détection des hélitrons entiers (composés des deux termini) et des hélitrons tronqués (une seule extrémité) avec notre modèle, aux séquences décrites par Kapitonov et Jurka [98]. Pour les hélitrons tronqués nous n'avons comparé que le nombre d'occurrences, car STAN ne détecte que la séquence terminale (Extrémité_5'_de_36pb :9 ou Extrémité_3'_de_36pb :9), et RepeatMasker détecte, en plus des 36 pb de cette extrémité, (une partie de) la séquence interne. Pour les hélitrons entiers, nous avons opposé, en plus du nombre d'occurrences, la taille moyenne des séquences avec la taille des consensus de Repbase. Nous avons considéré qu'une séquence détectée avait la "même" taille que le consensus Repbase si les deux tailles étaient égales à + ou - 10 % en pb.

Aussi bien pour les hélitrons tronqués que les entiers, les occurrences de toutes les familles d'hélitrons sont plus nombreuses avec STAN que celles observées avec RepeatMasker (Figure 3.4). Excepté les familles AtREP1, 5, 10D et 13, toutes les autres familles (séquences entières) ont plus ou autant de séquences ayant une taille similaire à la taille du consensus de Repbase. Cette valeur montre que STAN détecte plus d'hélitrons homologues au consensus que RepeatMasker. Ce dernier détecte principalement des séquences



Famille	Occurrences dans l'article	Taille du consensus dans Repbase (bp)	Occurrences détectées par RepeatMasker						Occurrences détectées par STAN			
			Sans terminus	5' termini	3' termini	Deux termini	Taille moyenne	Séquences proches de la taille du consensus	5' termini	3' termini	Deux termini	Taille moyenne avec les deux termini	Séquences proches de la taille du consensus
Helitron1	7	15809	155	6	3	0	706	0	84	55	12	7741	2
Helitron2	6	11435	135	2	9	0	915	2	29	21	15	6867	4
Helitron3	2	15333	39	3	2	0	1033	1	44	144	17	4507	1
Helitron4	5	17261	102	5	5	0	1260	0	14	14	3	8683	1
Helitron5	X	12495	65	2	5	0	749	1	482	449	325	1834	4
HelitronY1A	10	1348	72	11	29	2	377	6	86	429	41	2385	10
HelitronY1B	10	1311	61	11	10	0	441	8	79	472	52	2879	13
HelitronY1C	10	3058	56	0	6	0	329	2	189	336	68	2849	5
HelitronY1D	X	2541	102	14	15	1	314	2	45	228	22	4371	5
HelitronY1E	X	1291	43	3	7	0	347	3	21	566	9	1890	5
HelitronY2	10	11114	110	6	2	0	775	2	32	25	17	6605	4
HelitronY3	X	5166	214	16	4	0	380	0	38	11	9	4478	1
HelitronY3A	X	3315	26	3	0	0	226	1	47	86	17	4582	2
AtREP1	100	888	70	49	42	0	562	74	466	547	375	1869	69
AtREP2	150	564	40	65	57	0	463	113	488	570	379	1967	154
AtREP2A	X	603	28	38	31	1	413	56	488	577	392	1959	134
AtREP3	150	2097	300	86	93	13	649	51	220	672	196	2334	58
AtREP4	50	2240	118	21	24	0	454	7	60	70	48	3376	18
AtREP5	50	2386	89	18	27	1	773	22	140	597	105	1968	21
AtREP6	30	1189	49	12	8	0	460	14	241	126	106	1971	17
AtREP7	50	940	43	31	14	0	473	25	231	88	77	2055	26
AtREP8	40	1077	35	16	11	0	566	17	216	115	85	2111	24
AtREP9	10	899	8	6	5	0	470	8	231	95	80	2096	27
AtREP10	50	899	48	16	11	0	373	16	250	82	98	2270	26
AtREP10A	20	1380	115	18	14	0	287	5	255	120	120	2510	17
AtREP10B	20	1821	86	24	2	0	363	2	255	117	121	2333	8
AtREP10C	20	653	110	30	10	0	266	9	241	159	135	2842	7
AtREP10D	X	777	81	21	16	0	288	17	264	159	135	2941	13
AtREP11	50	1053	121	35	38	0	458	32	453	633	385	1968	45
AtREP11A	X	1003	31	14	5	0	425	16	474	580	353	2132	42
AtREP12	10	1342	19	8	3	0	418	7	35	8	7	1324	6
AtREP13	30	648	77	48	25	0	322	42	231	74	62	2000	34
AtREP14	X	737	30	7	22	0	317	10	474	232	184	1656	19
AtREP15	X	1769	46	18	5	0	413	7	35	71	20	2362	10
AtREP20	X	2145	X	X	X	X	X	X	212	64	52	2666	17
AtREP21	X	484	X	X	X	X	X	X	82	189	53	1269	12
AtREPX1	20	2432	159	8	10	0	334	1	20	20	20	3798	4
Total	910	131503	2883	671	570	18	17399	579	7252	8801	4195	113448	865

FIG. 3.4 – Comparaison de l'identification des hélitrons entre RepeatMasker et notre méthode syntaxique. Le symbole X correspond à l'absence de résultats. La seconde colonne correspond au nombre d'occurrences de chaque famille décrit dans l'article de Kapitonov et Jurka [98]. La troisième colonne est la taille des consensus présent dans Repbase [94]. La taille moyenne des hélitrons est calculée avec l'ensemble des séquences détectées pour RepeatMasker et seulement l'ensemble des séquences détectés avec un modèle à deux extrémités pour STAN.

d'hélitrons plus petites que les séquences consensus. Ce résultat prouve ainsi que RepeatMasker ne détecte pas des hélitrons entiers mais plutôt des petites fractions d'hélitrons

(en général sans les extrémités). Au contraire, la taille moyenne des séquences détectées par STAN, pour une famille donnée, est plus grande que la taille du consensus (Figure 3.4). Cette différence provient de la détection de certains hélitrons qui incluent dans leurs séquences internes d'autres transposons ou hélitrons.

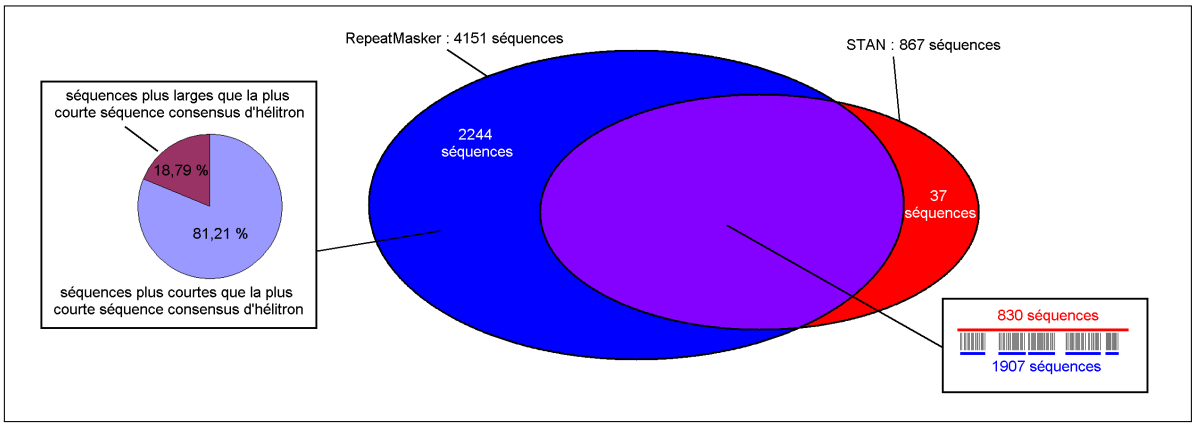


FIG. 3.5 – Comparaison des séquences détectées par STAN et RepeatMasker. Les séquences trouvées par les deux logiciels sont de couleur violette. Les séquences détectées seulement par RepeatMasker sont en bleu et celles détectées seulement par STAN sont en rouge. Le rectangle droit donne le nombre de séquences qui ont leurs positions communes pour RepeatMasker (en bleu) et STAN (en rouge). Comme une séquence détectée par une méthode donnée peut correspondre à plusieurs séquences de l'autre méthode (dans cet exemple, une séquence détectée par STAN est superposée avec 5 séquences détectées par RepeatMasker), les valeurs des séquences communes obtenues par les deux méthodes sont différentes. Le rectangle gauche montre le pourcentage de séquences détectées par RepeatMasker dont la taille est inférieure au plus petit consensus d'hélitron (564 bp) [98].

Nous avons aussi comparé l'ensemble des hélitrons détectés par les deux méthodes (Figure 3.5). Excepté pour 37 séquences, tous les hélitrons détectés par STAN sont aussi détectés par RepeatMasker. Au contraire, la plupart des séquences détectées par RepeatMasker n'ont pas été obtenues avec STAN. Plus de 80 % des séquences détectées par RepeatMasker ont été plus courtes que le plus court des hélitrons consensus (Figure 3.5) [98]. Ce résultat a confirmé que RepeatMasker détecte préférentiellement des fractions d'hélitrons.

3.2.3 Modèle hélitronique sans structure secondaire

Dans notre modèle syntaxique précédent, la structure secondaire de la tige-boucle subterminale du modèle biologique [98] a été remplacée par sa séquence primaire. Nous avons cherché à savoir si ce dernier modèle ne pouvait pas être amélioré en incluant la structure secondaire de la tige-boucle. La présence de la tige-boucle a donc été vérifiée dans toutes les familles d'hélitrons détectées par STAN [154] avec un modèle syntaxique.

Deux grammaires SVGs ont été écrites pour chaque famille (chaque séquence hélitronique de Repbase). Ces grammaires ont pris en compte la séquence palindromique de 6 à 8 nucléotides et la boucle contenue dans le palindrome. Le mot "*loop*", qui représente la boucle de l'hairpin, a été remplacée par la séquence consensus de chaque famille, en

admettant une erreur de substitution possible et deux indels. La première grammaire recherchait le palindrome exact, la deuxième autorisait une erreur de substitution dans ce palindrome. Elles ont été écrites dans la syntaxe STAN :

```
X:[6,8]-x(0,1)-loop:1-x(0,1)-~X ou X:[6,8]-x(0,1)-loop:1-x(0,1)-~X:1
```

Selon la famille d'hélitrons, 35 à 55 % des séquences hélitroniques ne présentent pas d'erreur de substitution dans la séquence palindromique de la tige boucle. Ce pourcentage passe de 55 à 80 % pour une tige-boucle ayant au maximum une erreur de substitution dans le palindrome. Bien qu'élevé, ce dernier pourcentage reste insuffisant pour considérer la tige-boucle comme un élément indispensable au fonctionnement de l'hélitron. Cette perte du palindrome parfait dans la tige-boucle est confortée par les résultats publiés par Laufs [122] démontrant qu'un géminivirus (même mode de réplication que l'hélitron) peut transposer faiblement sans hairpin. La plupart des hélitrons étant non-autonomes, nous pouvons penser que cette tige-boucle a subi un processus de dégénérescence au cours de son évolution qui ne permet plus la détection des hélitrons dans les génomes.

Résumé

Dans cette section, nous avons créé un modèle syntaxique capable de détecter l'ensemble des familles hélitroniques présentes dans le génome d'*Arabidopsis thaliana* (section 1.4). Cette détection par un modèle syntaxique est plus performante que la détection réalisée avec RepeatMasker, l'outil de référence de la détection des éléments transposables. Enfin, nous avons montré que le motif tige-boucle, motif caractéristique des hélitrons, n'est pas indispensable dans la détection des hélitrons.

Contribution scientifique

Article : Nicolas J., Durand P., Ranchy G., Tempel S. and Valin A.S. 2005. Suffix-Tree ANalyser (STAN) : looking for nucleotidic and peptidic patterns in genomes. Bioinformatics. 21 :4408-4410.

Conférence : Tempel S., Couée I., Nicolas J. and El Amrani A. 2004. Genome-wide analysis of domain organization in Arabidopsis helitronic structures. XII ème Colloque Eléments Transposables. Tours.

3.3 Analyse systématique et classification des hélitrons dans le génome d'Arabidopsis

Le modèle syntaxique des hélitrons sans hairpin que nous venons de définir nous a permis de réaliser une détection exhaustive des hélitrons chez *Arabidopsis thaliana*. Leur analyse permet de disposer de premiers indices sur le mode de transposition des hélitrons non-autonomes, l'évolution et le mécanisme d'invasion des différentes familles d'hélitrons.

3.3.1 Analyse des termini hélitroniques dans Arabidopsis

Pour une analyse exhaustive des hélitrons et de leurs extrémités, nous avons tout d'abord étudié séparément les extrémités hélitroniques, puis nous avons analysé les combinaisons 5'-3' des termini. Nous avons appelé "$left$" ("$right$"), l'extrémité 5' (3') d'une famille donnée d'hélitrons. LEFT et RIGHT sont respectivement définis comme l'ensemble des extrémités 5' et 3' de toutes les familles d'hélitrons connues dans Repbase [94].

3.3.1.1 Création de la matrice d'occurrences des hélitrons

Nous notons $left_i$ ($right_j$), l'extrémité 5' i (3' j) contenue dans l'ensemble des extrémités 5' (3'), nommé $LEFT$ ($RIGHT$). Pour chaque paire possible de termini ($left_i$, $right_j$) $\in LEFTXRIGHT$, nous avons créé une grammaire. Chaque grammaire a ensuite été utilisée pour l'analyse du génome d'*Arabidopsis thaliana* via le logiciel STAN. Le résultat de ces analyses a été reporté dans une matrice d'occurrences sur $LEFTXRIGHT$. Une cellule d'indice ($left_i$, $right_j$) contient le nombre d'occurrences du modèle constitué par l'extrémité 5' $left_i$ associée à un terminus 3' $right_j$ et séparé par un certain nombre de nucléotides. Cette définition autorise les chevauchements, l'insertion d'hélitrons et les chimères d'hélitrons. Une chimère est une séquence créée par la combinaison de deux séquences distinctes.

3.3.1.2 Agrégation des extrémités et des couples d'extrémités

Les tailles de $LEFT$ et $RIGHT$ sont élevées, ce qui signifie que les motifs des extrémités ont été très finement distingués. Dans le but de rationaliser le choix des termini, nous avons d'abord eu besoin de réduire ce nombre d'extrémités par la formation de classes d'équivalence.

Nous avons noté f^{ij}_{left} (f^{ij}_{right}), le nombre d'occurrences du terminus couvert par le motif $left_i$ ($right_i$) et non couvert par le motif $left_j$ ($right_j$). Par exemple, $left_2$ représente l'extrémité 5' de l'AtREP2 et $left_{14}$ l'extrémité 5' de l'AtREP14. Les deux extrémités ont seulement une substitution de différence ; en les recherchant avec STAN et neuf erreurs, il est normal de retrouver des positions communes aux deux détections. Néanmoins, certaines positions ne seront détectées que par AtREP2 : ces positions représentent dans ce cas $f^{2\ 14}_{left}$ (Figure 3.6).

Nous avons appliqué sur l'ensemble des termini un algorithme standard de classification hiérarchique, commençant par l'ensemble des singletons qui correspondaient à l'ensemble des termini. L'algorithme agrège à chaque étape les classes qui sont à une distance minimale. La distance entre deux classes c_1 et c_2 est définie comme suit :

$d(c_1, c_2) = Min_{x \in (c_1 \cup c_2)} \sum_{y \in (c_1 \cup c_2) - x} f^{yx}$

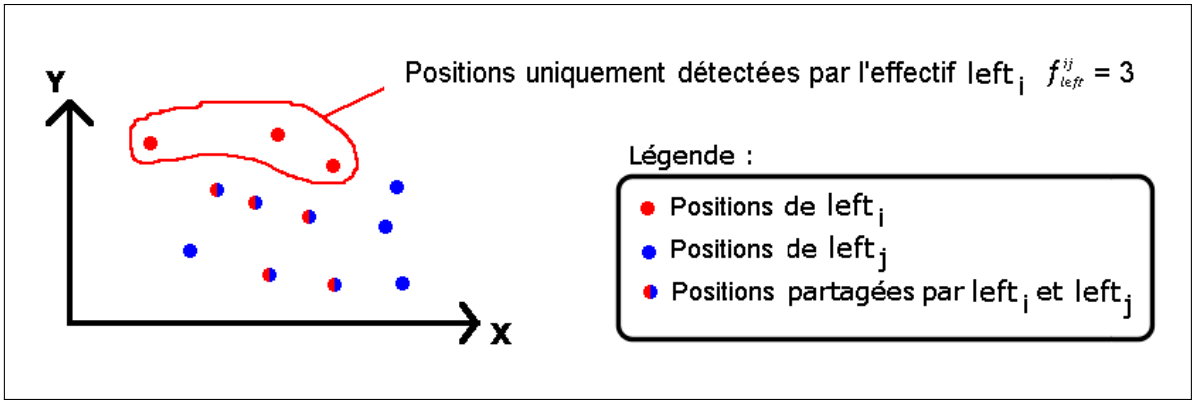


FIG. 3.6 – Visualisation d'occurrences i non couvertes par j et exemple d'une distance entre deux séquences. La première partie de la figure montre les positions de $left_i$ et $left_j$. Les positions de $left_i$ non couvertes par $left_j$ sont entourées en rouge.

Informellement, la distance entre les deux classes c_1 et c_2 est le "rayon" de la classe agrégée obtenu en minimisant le nombre d'occurrences non couverts par un élément de cette classe (élément qui devient de fait un représentant de cette classe). Nous avons retenu l'agrégation de deux classes si la distance obtenue est inférieure à 10 % du total du nombre d'instances couvertes par $c_1 \cup c_2$. A partir de cette union, le représentant de la nouvelle classe est l'élément qui minimise la taille du rayon obtenue par l'agrégation des deux classes (i.e. c_1 si le nombre d'occurrences perdues par ce choix est inférieur au nombre d'occurrences perdues par le choix de la classe c_2). La nouvelle classe est étiquetée par son représentant. Dans notre exemple précédent, AtREP2 a 488 occurrences dans le génome et AtREP14 a 478 occurrences. Ces deux classes ont 478 occurrences en commun et leur distance est inférieure à 10 % . La nouvelle classe est étiquetée par AtREP2.

A partir de la matrice brute d'occurrences obtenues par l'ensemble des combinaisons de termini, nous avons utilisé deux méthodes pour classer les éléments de cette matrice en fonction de la similarité des valeurs des cellules : l'algorithme d'optimisation combinatoire de Munkres [152] et une classification manuelle.

L'algorithme de Munkres s'attaque au problème de l'affectation des tâches, c'est-à-dire à la minimisation de la somme des coûts des tâches à effectuer pour un ensemble d'actions [152]. Dans notre cas, les acteurs sont représentés par les extrémités 5', les tâches par les termini 3' et le coût des tâches est représenté par le nombre d'occurrences non couverts par le choix d'affectation de 3' à 5'. Autrement dit, à chaque cellule ij (ligne i, colonne j) le coût représente la somme des occurrences de toutes les autres cellules de la ligne i moins l'occurrence de cette cellule ij ; plus formellement, le coût est égale à $c(ij) = \sum occurrences(i.) - occurrences(ij)$. L'algorithme trouve, pour l'ensemble des extrémités 5', le meilleur choix d'extrémité qui minimise la somme des occurrences détectées. Nous avons ensuite appliqué l'algorithme à la nouvelle matrice : pour chaque extrémité 5' donnée, l'algorithme choisit le terminus 3' qui ne sera pas forcément celui qui donne la plus petite valeur d'occurrence pour cette extrémité 5' donnée, mais celui qui minimise la somme des occurrences obtenues pour l'ensemble des extrémités 5' associées avec une extrémité 3'. Chaque extrémité 3' n'est associée qu'avec un seul terminus.

Après cet algorithme, nous avons obtenu une matrice optimisée où une extrémité 5' donnée est associée avec une extrémité 3' donnée. Néanmoins, les autres combinaisons de chaque extrémité 5' (3') qui n'ont pas été sélectionnées par l'algorithme, ne sont pas classés en fonction de leur similarité. Nous avons donc réalisé la réorganisation de la matrice en fonction de ces autres combinaisons, en conservant pour chaque ligne (extrémité 5') et colonne (extrémité 3'), l'association choisie par l'algorithme d'optimisation combinatoire.

3.3.1.3 Analyse des occurrences des termini 5' et 3' : évidence de la présence d'hélitrons tronqués

Nouveau nom	Nom de la famille d'hélitron	Séquences	Occurrences	Perte (%)
1	HELITRON1	TCTACATATACATTTTTGGGGACGATTTGTGTCGAA	84	0,00
2	HELITRON2, Y2 (HelitronY2)	TCTACACTATTATTTGAGACGTACGTTAAGTGATTC	32	9,06
3	HELITRON3	TCTACTTAACATTTTTGAAGTACAAAATAAGGAATT	44	0,00
4	HELITRON4	TCTATACTATTAAGAGAGAATGAGGGACAGAAAATT	14	0,00
5	AtREP1, 11 (AtREP11)	TCTACATATACATTTTTGCAGCCATTTTAGCAAATA	453	2,43
6	Helitron5, AtREP2, 2A, 11A, 14 (AtREP2)	TCTATATATACATTTTTGCAGCCATTTTGTGAAATA	488	5,33
7	HELITRONY1A	TCTACATATACATTTTTGGTAGGGTTTTTGGCCAAA	86	0,00
8	HELITRONY1B	TCTACATATACATTTTTGGTAGACTTTTTGGCCAAA	79	0,00
9	HELITRONY1C	TCTACATATACATTTTTGGAACGATTTGTGTAGAAA	189	0,00
10	HELITRONY1D	TCTACTTAACATTTTTGAAGTACATTTTAAGGAATC	45	0,00
11	HELITRONY1E	TCGACATTTTAAATCAAATATTTACCTAACAAAAAT	21	0,00
12	HELITRONY3	TCTTATATAATAAGAGAAGGTTTTTTCTAACTTTGC	38	0,00
13	HELITRONY3A	TCTACTTAACATTTTTGAAGTACAAAATAAGGATTT	47	0,00
14	AtREP3, 20 (AtREP3)	TCCTACTATATTATTTGGGAAGTACATTTTAAATGT	220	0,00
15	AtREP4	TCTATATTATTAATTGGGAAGCATTGTGTGAACTAC	60	0,00
16	AtREP5	TCTATATATACATTTTTGGAGGGGATTTTTGAAGAT	140	0,00
17	AtREP6, 7, 8, 9, 13 (AtREP6)	TCATATATATGAAAGTTGGCCAACTCTCTCCATATA	241	6,64
18	AtREP10, 10A, 10B (AtREP10A)	TCTTATTATATAAAGTATGGTTTTCAAAGTTACTAA	255	1,96
19	AtREP10C, 10D (AtREP10D)	TCTTATTATATAAAGTATGGTTTTTAAAATTACTAA	264	3,41
20	AtREP12	TCTAAATATACTAAATCAGCAGTCACTTTTCCAATA	35	0,00
21	AtREP15	TCTATACTATTAATTGGGAATCATTGAAAATAAACA	35	0,00
22	AtREP21	TCCCTTTATTATTAAAAGGGAAGTACAAATTGAAAT	82	0,00
23	AtREPX1	TCAACACCATAAAAACACTAAAAGTCTCTCCTGTGC	20	0,00
		Total	2972	2,39

FIG. 3.7 – Occurrences et regroupement des extrémités 5' dans le génome d'*Arabidopsis thaliana*. La première colonne est le nouveau nom (un chiffre) de l'extrémité 5'. La seconde colonne donne les anciens noms des extrémités, avec entre parenthèses le nom conservé pour le groupe d'extrémités de cette ligne. La troisième indique la séquence consensus conservée pour cette extrémité. La quatrième colonne indique le nombre d'occurrences du terminus conservé. Enfin, la dernière colonne donne le nombre d'occurrences perdues en choisissant cette extrémité pour représenter ce groupe d'extrémités.

L'analyse de la distribution de chaque terminus de chaque hélitron a montré que certaines extrémités sont préférentiellement regroupées avec des extrémités appartenant à d'autres hélitrons (Figure 3.7 et 3.8). Le regroupement des extrémités hélitroniques a réduit le nombre d'extrémités de 37 à 23 pour les extrémités 5' et 3'. Par exemple, les extrémités 5' et 3' des AtREP2 et 2A sont associées aux extrémités 5' et 3' des AtREP6, 7, 8, 9. Par contre, nous avons noté qu'un certain nombre d'extrémités ne sont jamais associées avec d'autres termini, tel que AtREPX1. Curieusement, des extrémités 5' de certaines familles d'hélitrons étaient associées alors que les extrémités 3' de ces mêmes familles n'étaient pas associées ou associées avec des extrémités d'autres familles. Par exemple, le terminus 5' d'AtREP3 est regroupé avec l'AtREP20 (Figure 3.7) et l'extrémité

3' est associée avec l'AtREP11 (Figure 3.8).

Curieusement, excepté les familles Helitron2, 4, Y2, AtREP4 et X1, le nombre d'occurrences d'un terminus 5' d'une famille donnée n'est pas similaire au nombre d'occurrences du terminus 3' correspondant. Ainsi, l'extrémité 5' de l'AtREP20 présente un nombre d'occurrences trois fois plus élevé que l'extrémité 3' correspondante (Figure 3.7 et 3.8). De même, l'extrémité 3' de l'AtREP3 est trois fois plus fréquente que son extrémité 5'. L'analyse de ces termini 3' en excès a révélé que certains d'entre eux ne sont associés à aucune extrémité 5'. Ces termini 3' correspondent à des hélitrons tronqués.

Nouveau nom	Nom de la famille d'hélitron	Séquences	Occurrences	Perte (%)
a	HELITRON1, Y1C (HELITRONY1C)	TAATCAACCCGCGGTGTACCGAGGGTCAATATCTAG	336	2,38
b	HELITRON2, Y2 (HELITRONY2)	CCAAAGATACCGTGCGTAGCACGGGTACTGACCTAG	25	0,00
c	HELITRON3	AAAAAATCTCGCGGTGTACCACGGGTCATATCCTAG	144	0,00
d	HELITRON4	TTTAACACCCGGGCGAAGCACGGGTTATCAATCTAG	14	0,00
e	HELITRON5	AACTATCCCTGCGGTGTACCGAGGGTCAAAATCTAG	449	0,00
f	HELITRONY1A	ATATCCACCCGCGGTACACCGCGGGTCAATATCTAG	429	0,00
g	HELITRONY1B	ATATCAACCCGCGGTGTACCGCGGGTCAATCTCTAG	472	0,00
h	HELITRONY1D	TATTTAACCCGCGGTATACCGCGGGTCATATCCTAG	228	0,00
i	HelitronY1E, AtREP11A (AtREP11A)	ATATAGCCCCGCGGTATACCGCGGGTTAAATTCTAG	580	7,59
j	HELITRONY3	TATATAAAATCCACGCATCGCGTGGACAACTTCTAG	11	0,00
k	HELITRONY3A	CAAAAATCCTGCGGTGTACCGCGGGTCATATCCTAG	86	0,00
l	AtREP1, 2, 2A (AtREP2A)	AATTGTCTCCGCGGTATACCGCGGGTTAAAATCTAG	577	7,63
m	AtREP3, 11 (AtREP3)	AAATCGTCCCGCGGTATACCGCGGGTTAAAATCTAG	672	5,51
n	AtREP4, 15 (AtREP15)	ATGAAATCCCGCACGTACGTGCGGGTCAGGATCTAG	71	2,82
o	AtREP5	ATATAGGCCCGCGGTATACCGCGGGTTAATATCTAG	597	0,00
p	AtREP6, 7, 8, 9 (AtREP6)	AAAACAAACCCACGCGTAGCGTGGGTACTCATCTAG	126	7,14
q	AtREP10, 10A, 10B, 10C, 10D (AtREP10C)	ATCTTACACCCGCCTATTTAGGCGGGCCTTATCTAG	159	6,29
r	AtREP12	CAATGTTGTCCCTGCATAGCAGGGGCCAAATGCTAG	8	0,00
s	AtREP13	CAAAAAAACCGCGGCGTAGCGCGGGTGATCACCTAG	74	0,00
t	AtREP14	TATTTGCCCCGTGGTGTACCACGGGTTAAATCCTAG	232	0,00
u	AtREP20	AAATGCCCGTGCTGTATAGCACGGGTTATGATCTAG	64	0,00
v	AtREP21	CCGATTGCCCGCGGTAAACCGCGGGTTAAAACCTAG	189	0,00
w	AtREPX1	AAAAAAGCCCGCGGCTTGGCGCGGGTTTCGCCCTAG	20	0,00
		Total	5563	2,77

FIG. 3.8 – Occurrences et regroupement des extrémités 3' dans le génome d'*Arabidopsis thaliana*. La première colonne est le nouveau nom de l'extrémité 3'. La seconde colonne donne les anciens noms des extrémités, avec entre parenthèses le nom conservé pour le groupe d'extrémités de cette ligne. La troisième indique la séquence consensus conservée pour cette extrémité. La quatrième colonne indique le nombre d'occurrences du terminus conservé. Enfin, la dernière colonne donne le nombre d'occurrences perdues en choisissant cette extrémité pour représenter ce groupe d'extrémités.

3.3.1.4 Distribution des combinaisons d'extrémités hélitroniques

Toutes les paires possibles de termini 5' et 3', en accord avec le modèle de la Figure 3.3, ont été recherchées avec STAN [154] dans le génome entier d'*Arabidopsis thaliana*. Nous avons appliqué l'algorithme d'optimisation combinatoire sur cette matrice, puis réalisé la réorganisation manuelle. Le nombre d'occurrences de chaque paire est donné dans la Figure 3.9 et l'ensemble des positions de chaque paire est rassemblé dans le tableau de l'annexe 5.1. Toutes les familles d'hélitrons précédemment connues [98, 94] ont été retrouvées avec le nombre attendu d'occurrences. De plus, l'alignement multiple des séquences obtenues d'une combinaison 5'- 3' d'extrémités d'une famille donnée avec son consensus a montré que ces occurrences sont similaires au consensus de Repbase.

Un grand nombre de nouvelles occurrences apparaît par rapport à l'état de l'art. Ce résultat augmente le nombre estimé d'hélitrons chez Arabidopsis de 870 copies [98] à 1504 copies (Figure 3.5). Ces nouvelles occurrences correspondent à des combinaisons de termini non détectées précédemment : par exemple, le nombre de 5' terminus nouvellement nommé 14 est fréquemment associé avec le terminus 3' nommé "l" (171 occurrences Figure 3.9).

5' \ 3'	s	n	p	q	t	o	l	m	e	i	f	g	a	v	h	c	k	u	j	b	r	d	w
21	2	20	2	4	1	4	2	5	1	4	4	1	1	1	1	0	0	0	0	0	0	0	0
15	2	47	0	2	1	13	12	19	4	10	9	9	6	1	1	3	2	1	0	0	1	0	0
17	62	4	106	9	13	18	20	29	17	21	13	21	14	10	15	2	15	4	0	1	0	0	1
18	3	8	5	136	13	23	24	27	21	25	23	16	8	9	9	2	5	2	0	1	0	2	1
19	3	10	5	135	14	24	26	29	23	26	24	17	9	9	9	3	5	3	0	1	0	2	1
1	1	0	2	0	21	57	58	62	41	54	43	46	32	11	23	9	7	3	0	1	0	0	0
5	9	3	5	6	181	309	369	378	317	336	150	165	133	79	119	29	30	14	1	1	1	1	2
6	8	6	5	6	182	335	392	402	323	362	175	189	149	87	119	27	27	14	1	3	1	1	1
9	6	2	4	4	64	120	129	141	115	127	69	91	68	21	42	19	17	2	1	1	1	0	1
16	2	3	4	3	32	105	111	116	88	106	71	68	54	32	36	4	1	4	0	0	0	0	1
14	0	9	1	6	48	163	171	196	117	151	150	137	134	64	70	28	27	52	2	0	0	1	0
7	1	0	1	2	5	38	23	33	23	29	41	55	34	5	11	5	3	0	0	0	0	0	2
8	2	0	1	4	11	34	21	34	22	30	34	52	32	2	13	5	6	0	0	0	0	1	2
22	0	2	0	3	3	19	19	21	6	18	10	12	9	53	13	3	3	0	0	0	0	0	0
3	3	0	2	0	12	10	7	15	7	12	5	12	13	2	24	17	17	0	0	0	0	1	0
10	2	0	2	0	10	7	5	12	5	9	4	11	11	1	22	17	16	0	0	0	0	1	0
13	3	0	3	0	12	9	7	14	7	11	5	12	13	2	23	17	17	0	0	0	0	1	0
11	0	0	0	0	5	9	9	9	8	9	7	8	4	1	2	1	1	1	0	0	0	0	0
12	0	0	0	0	2	5	3	4	3	4	3	5	5	0	1	0	1	0	9	0	0	0	0
2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	17	0	0	0
20	0	0	0	0	1	4	2	4	2	2	1	1	2	0	1	1	0	1	0	0	7	0	0
4	0	0	0	0	2	1	2	1	1	1	1	1	1	0	1	0	2	0	0	0	0	3	0
23	4	1	2	1	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0	1	0	0	14

x<10 x<20 x<30 x<50 x<100 x<200 x>200

FIG. 3.9 – Matrice de fréquence des modèles associés aux paires d'extrémités d'hélitron. Les lignes et les colonnes ont été classées par l'algorithme d'optimisation de Munkres [152]. Chaque ligne représente un terminus 5' de 36 pb et chaque colonne représente une extrémité 3' de 36 pb.

Une nouvelle comparaison des séquences hélitroniques a été réalisée entre les séquences de RepeatMasker et les séquences de la matrice d'occurrences découvertes avec STAN. A notre surprise, peu de nouvelles occurrences correspondent à de nouvelles séquences. Certaines nouvelles paires montrent un nombre significatif de copies dans le génome, ce qui semble indiquer que ces nouvelles combinaisons ont effectivement fait l'objet de transpositions et peuvent être considérées comme de nouveaux hélitrons. Les séquences internes de ces combinaisons constituent des combinaisons de séquences internes d'hélitrons connues et de nouvelles séquences. Les nouvelles combinaisons présentent en majorité les mêmes séquences (mêmes positions) que les 37 familles d'origine. Certaines occurrences de nouvelles combinaisons ont été détectées à des positions identiques à celles d'autres occurrences précédemment connues. A cause de nombreuses mutations, les extrémités à ces positions possèdent des motifs caractéristiques de deux extrémités différentes. Ce résultat suggère un lien évolutif entre ces combinaisons (nouvelles et connues). Les extrémités ont pu muter d'une extrémité consensus connue pour créer une autre extrémité consensus

connue. Les extrémités observées à ces positions seraient des intermédiaires entres les deux extrémités consensus. D'autres occurrences ont des positions différentes pour une même séquence, autrement dit la séquence hélitronique est flanquée de plusieurs combinaisons d'extrémités. L'ensemble des résultats a confirmé qu'une séquence interne d'hélitron n'est pas spécifique d'une famille de Repbase donnée.

Les termini 5' et 3' d'une famille donnée sont généralement hautement corrélés. Par exemple, la principale association du terminus 5' de l'AtREP4 est avec le terminus 3' de ce même hélitron (combinaison 15-n dans la Figure 3.9). Néanmoins, chaque extrémité a montré un niveau significatif d'association avec des termini de certaines autres familles, alors que certaines associations telles que la paire 16-j ne donne aucune occurrence. De plus, le pattern de distribution des associations n'est pas aléatoire, mais est clairement concentré en clusters d'associations. Ainsi, les termini 5' numéro 18 et 19 sont beaucoup plus associés avec le même terminus 3' nommé q (Figure 3.9, 3.7 et 3.8). Finalement, quelques termini montrent plus d'interactions avec les termini d'autres familles qu'avec l'autre extrémité de leur propre famille. Par exemple, le terminus m (3' d'AtREP3) est associé 378 fois avec l'extrémité 5 (5' d'AtREP1) et seulement 196 fois avec son propre terminus 5' (Figure 3.9).

3.3.2 Analyse du mode de transposition des hélitrons

A partir de l'étude précédente des occurrences des termini, nous avons pu détecter toutes les familles potentielles d'hélitrons. Nous nous sommes donc intéressés au mode de réplication de ces éléments dans ce génome. Nous avons analysé les occurrences des combinaisons hélitroniques : les plus répandues dans le génome devraient être les extrémités qui présentent le plus d'affinité avec les protéines de transposition codées par les hélitrons autonomes.

Nous avons commencé à agréger les termini grâce à une méthode similaire à celle qui a été utilisée pour les combinaisons d'extrémités (section 3.3.1.2). Dans ce cas, f^{ij}_{left} (f^{ij}_{right}) représentait toutes les positions obtenues des hélitrons entiers qui possèdent l'extrémité 5' (3') i et qui ne sont pas obtenues par des hélitrons entiers avec l'extrémité 5' (3') j. Pour les deux extrémités 5' (3') i et j, nous comparons l'ensemble de leurs combinaisons avec tous les termini RIGHT (termini LEFT). Nous avons noté la classe c_1 l'ensemble des positions d'une extrémité hélitronique et la classe c_2 l'ensemble des postions détectées avec une autre extrémité. Nous avons agrégé c_1 et c_2 si le nombre d'occurrences qui ne sont pas contenues dans $c_1 \cap c_2$ est inférieur à 10 % d'occurrences obtenues par $c_1 \cup c_2$.

Après l'agrégation des termini 5' et 3', nous avons réalisé une réorganisation manuelle pour séparer la matrice en clusters, en fonction des valeurs des cellules.

3.3.2.1 Les clusters d'occurrences suggèrent différentes activités de transposition

La Figure 3.10 montre la structure de la matrice des occurrences observées après la réorganisation manuelle des lignes et des colonnes. Quatre clusters d'occurrences ont pu être déduits de cette matrice.

Le premier cluster (rectangle bleu en haut à gauche de la Figure 3.10) correspond principalement à une suite de familles d'hélitrons similaires à ceux décrit dans Repbase, mais

de fréquence relativement faible. Seules quelques combinaisons de ce cluster présentent un nombre notable d'occurrences.

Le second cluster (en haut à droite) est composé de nouvelles paires de termini. Par exemple, la nouvelle combinaison de l'extrémité 5' numéro 18-19 avec le terminus 3' noté o,i,l,m n'était pas décrite dans Repbase et présente un grand nombre d'occurrences (Figure 3.10).

5' \ 3'	j	w	s	p	b	n	q	r	d	u	t	v	e	o,i,l,m	f	g	a	h	k	c
12	9	0	0	0	0	0	0	0	0	0	2	0	3	3	3	5	5	1	1	0
23	0	14	4	2	1	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0
17	0	1	62	106	1	4	9	0	0	4	13	10	17	26	13	21	14	15	15	2
2	0	0	0	1	17	1	1	0	0	1	1	1	1	1	1	1	1	1	0	0
15	0	0	2	0	0	47	2	1	0	1	1	1	4	14	9	9	6	1	2	3
21	0	0	2	2	0	20	4	0	0	0	1	1	1	3	4	1	1	1	0	0
18,19	0	1	3	5	1	8	136	0	2	2	13	9	21	27	23	16	8	9	5	1
20	0	0	0	0	0	0	0	7	0	1	1	0	2	2	1	1	2	1	0	1
4	0	0	0	0	0	0	0	0	3	0	2	0	1	1	1	1	1	1	2	0
22	0	0	0	0	0	2	3	0	0	0	3	53	6	29	10	12	9	13	3	3
14	2	0	0	1	0	9	6	0	1	52	48	64	117	204	150	137	134	70	27	28
9	1	1	6	4	1	2	4	1	0	2	64	21	115	136	69	91	68	42	17	19
5,6	1	1	8	5	3	6	6	1	1	14	182	87	323	402	175	189	149	119	27	27
16	0	1	2	4	0	3	3	0	0	4	32	32	88	111	71	68	54	36	1	4
1	0	0	1	2	1	0	0	0	0	3	21	11	41	56	43	46	32	23	7	9
7	0	2	1	1	0	0	2	0	0	0	5	5	23	28	41	55	34	11	3	5
8	0	2	2	1	0	0	4	0	1	0	11	2	22	26	34	52	32	13	6	5
3,10,13	0	0	3	2	0	0	0	0	1	0	12	2	7	10	7	12	13	24	17	17
11	0	0	0	0	0	0	0	0	0	1	5	1	8	9	7	8	4	2	1	1

x<10	x<20	x<30	x<50	x<100	x<200	x>200

FIG. 3.10 – Matrice de fréquence des occurrences de toutes les paires possibles 5' - 3' correspondant au modèle de la Figure 1.19. Chaque cellule est coloriée en fonction de sa fréquence. Chaque ligne représente un terminus 5' ou une agrégation de termini 5' de 36 pb et chaque colonne représente une extrémité 3' ou une agrégation d'extrémités 5'. Les rectangles bleus délimitent les quatre clusters de familles.

Le troisième cluster (en bas à gauche) est caractérisé par un très faible nombre d'occurrences pour ces combinaisons de termini. Nous pouvons raisonnablement suggérer que ces combinaisons ne sont pas favorables à la transposition ou que l'association n'est pas récente, si on y découvre un hélitron autonome.

Enfin, le dernier cluster (en bas à droite) correspond à la majorité des occurrences d'hélitrons présents dans le génome complet d'*Arabidopsis thaliana* (Figure 3.10).

Cette différence d'occurrences peut être liée à une différence de reconnaissance des termini par les protéines de transposition des hélitrons (RPA + hélicase). On peut supposer que les protéines de transposition reconnaissent préférentiellement les termini identiques ou similaires aux termini des hélitrons dont elles étaient issues [70]. Pour vérifier cette hypothèse, nous avons recherché les hélitrons autonomes présents dans cette matrice d'occurrences.

3.3.3 Identification de nouvelles familles d'hélitrons autonomes et non-autonomes

Puisque les hélitrons autonomes sont normalement requis pour la transposition des hélitrons, autonomes et non-autonomes [70], nous avons cherché à identifier l'ensemble des hélitrons contenant des ORFs et susceptibles d'être autonome. Pour cela, nous avons utilisé GENSCAN (genes.mit.edu/GENSCAN.html) [26] afin qu'il détecte tous les ORFs putatifs d'hélitrons et ensuite BLASTP [3] pour qu'il identifie tous ces ORFs (www.ncbi.nlm.nih.-gov/BLAST/).

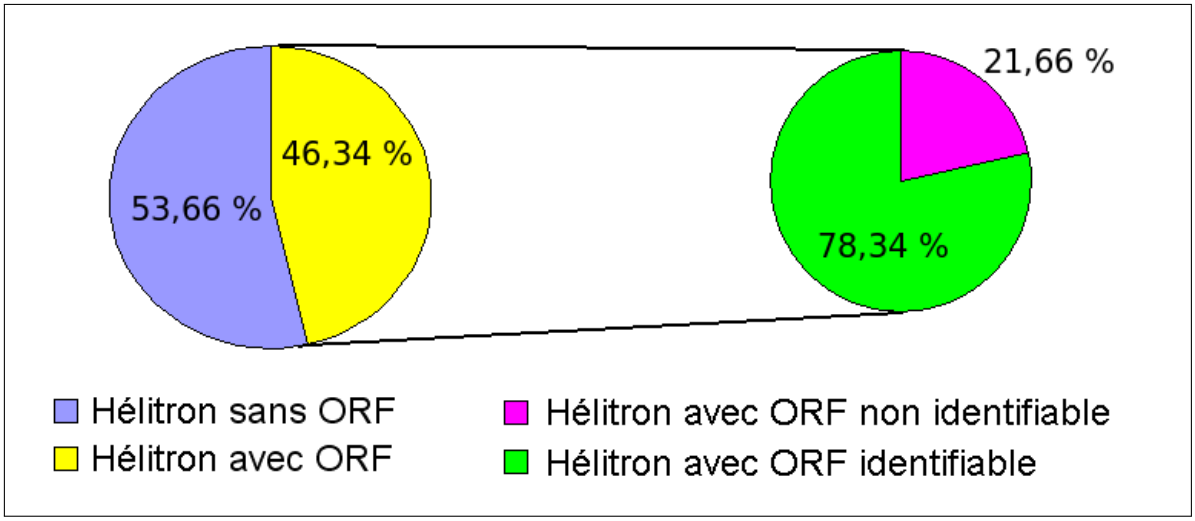


FIG. 3.11 – Pourcentage d'hélitrons possédant un ORF dans leur séquence et proportion de ces ORFs identifiés par BLASTP.

La majorité des hélitrons (807 sur 1504) ne contient pas d'ORFs dans leur séquence interne (Figure 3.11). Parmi les 697 séquences qui sont composées d'ORFs, 151 contiennent des petits fragments d'ORF reconnus par GENSCAN, mais non reconnus par BLASTP. Cette non-reconnaissance peut provenir de la taille des ORFs (inférieure à 20 AA), ou de l'absence d'homologue dans la banque de données du NCBI. Néanmoins, l'alignement de ces ORFs non reconnus a montré de grands groupes d'ORFs très différents les uns des autres. Les séquences de chaque groupe sont très similaires. Parmi les 546 ORFs s'appariant avec un candidat proposé par BLASTP, une partie d'entre eux (144) le fait avec des protéines étiquetées "inconnues" ou "hypothétiques". Leur alignement ne montre aucune similarité entre elles.

Il est possible que ces ORFs soient des portions de pseudogènes [75]. Par ailleurs, la plupart des ORFs détectés sont des gènes présents en un seul exemplaire dans l'ensemble des hélitrons. Ce résultat suggère que ce n'est pas le gène codant une protéine qui est inséré dans un hélitron, mais plutôt que les deux extrémités hélitroniques se sont insérées séparément lors de deux événements de transposition. Ainsi, ces hélitrons ne seraient pas des hélitrons "entiers" (composé d'une extrémité 5' et 3'), mais des hélitrons tronqués. Néanmoins de nombreux exemples chez le maïs [149, 119, 210] montrent que la transposition de ces ORFs est possible. Par exemple, nous avons trouvé le même gène codant pour la protéine hypothétique nommée "F10O5.11" présente sur le chromosome 1 et 3

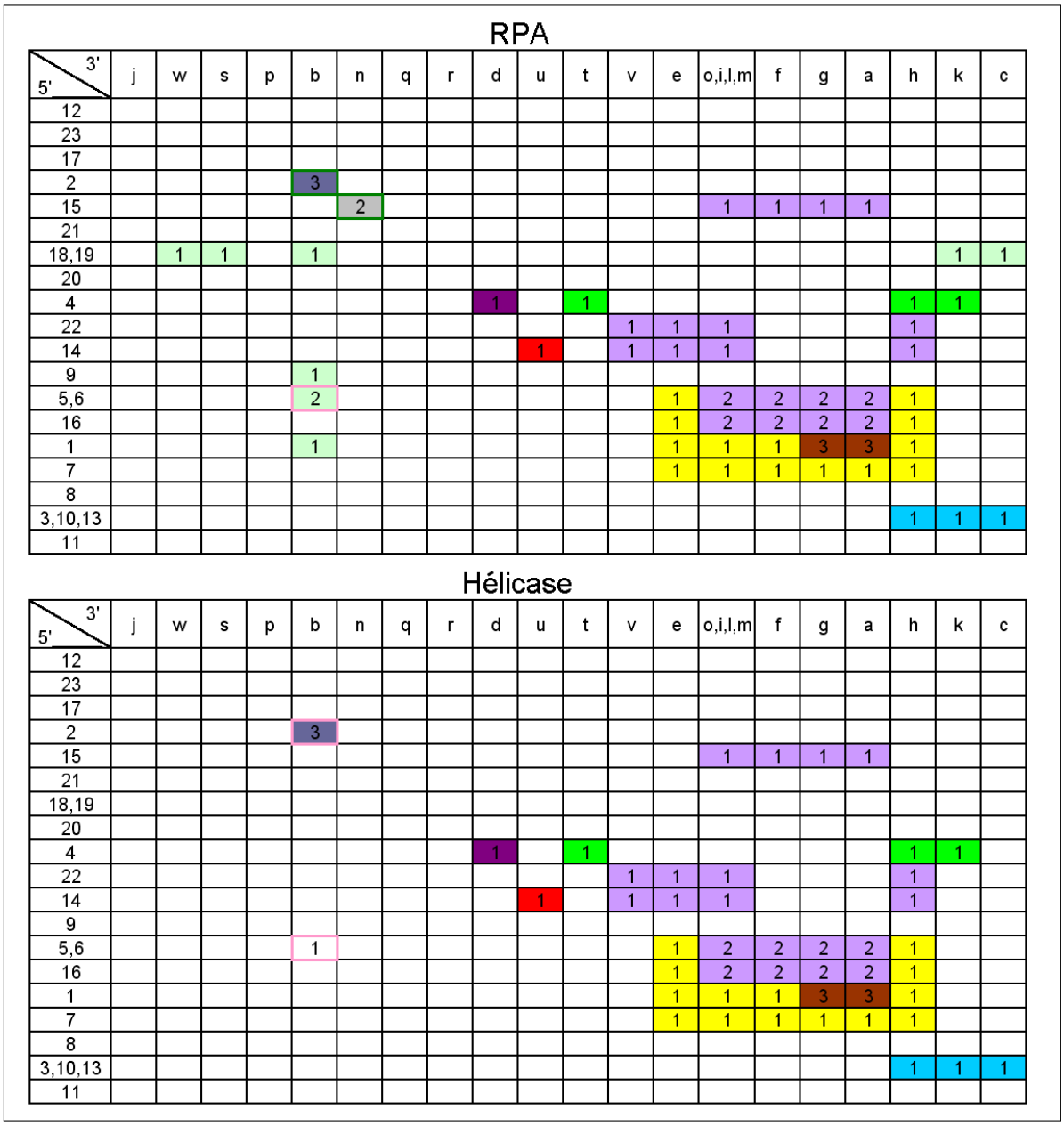


FIG. 3.12 – Occurrences des hélitrons codant pour des ORFs de RPA-like et d'hélicase-like. Chaque couleur donnée représente la même séquence hélitronique avec une ou plusieurs combinaisons de termini.

d'Arabidopsis. Ce résultat nous a suggéré que les protéines de transposition ont reconnu les deux extrémités intégrées distinctement et ont réalisé un événement de transposition des deux extrémités et de la séquence entre ces deux termini.

Nous avons trouvé 66 séquences hélitroniques contenant des ORFs codant pour des hélicase-like et/ou des RPA-like. Les protéines hélicase-like sont toujours associées avec les protéines RPA-like. Nous avons aussi trouvé que, curieusement, certaines paires de termini qui contiennent des hélicase-like et des RPA-like se retrouvent toujours associées à d'autres extrémités par simple effet de similarité. Ainsi, l'extrémité 5' numéro 16 est similaire à l'extrémité 5' numéro 1 à quelques substitutions près. L'analyseur STAN [154] a reconnu

cette extrémité hélitronique comme les deux termini 5' 1 et 16. Certains hélitrons semblent contenir des RPA-like (au moins 2 ORFs) sans hélicase-like (Figure 3.12), néanmoins ils contiennent toujours un autre ORF non identifié par BLASTP. Un alignement de cette "protéine inconnue" avec une hélicase a montré des similarités très dispersées mais significatives. Il semble probable que ces protéines non identifiées soient des hélicases qui aient subi de nombreuses mutations. Ainsi, la séquence hélitronique identifiée par les extrémités 5' 14 - u 3' et située sur le chromosome 1 (position 19728423 à 19740251) n'est pas reconnue comme une hélicase par BLASTP, mais l'alignement de la protéine prédite avec des hélicases, consensus décrite par Kapitonov et Jurka [98], montre une similarité presque parfaite sur une centaine d'acides aminés en 5' (la protéine est tronquée sur la partie centrale et la partie 3' (800 à 1000 AA)).

Curieusement, la majorité des séquences qui contiennent des protéines RPA-like (40 sur 44 occurrences) et des protéines hélicase-like (25 sur 32 occurrences) possèdent plusieurs paires d'extrémités hélitroniques (Figure 3.13). Par exemple, l'alignement multiple avec ClustalW [200] des différentes combinaisons, contenant les protéines de transposition de la position 3200666 à 3210806 dans le chromosome II, a montré qu'il s'agissait d'une seule séquence hélitronique qui possède plusieurs combinaisons de termini dans les deux sens de l'ADN (Figure 3.13). Les multiples combinaisons de termini pour une seule séquence d'hélitron (observées aussi pour les hélitrons non-autonomes) expliquent en partie la large différence entre le nombre d'occurrences des extrémités et les paires d'extrémités (Figure 3.7, 3.8 et 3.10).

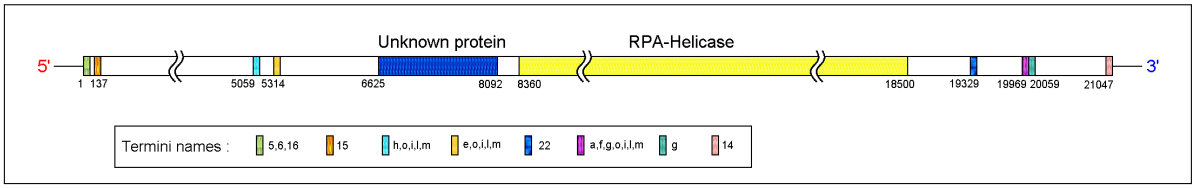


FIG. 3.13 – Visualisation de multiples termini autour d'un gène codant une seule RPA-hélicase protéine. La couleur rouge (bleue) montre les extrémités 5' (3'). Les deux ORFs n'ont pas la même orientation (protéine inconnue : orientation + ; RPA-hélicase : orientation -).

De façon intéressante, certains hélitrons autonomes contiennent des ORFs de "protéines inconnues" dans leur séquence. Néanmoins, ces ORFs sont de très petites tailles (inférieure à 100 AA) et n'ont donné aucun hit avec BLAST. Des ORFs supplémentaires ont aussi été découverts dans les hélitrons autonomes de la chauve-souris [169]. Ce résultat pourrait souligner l'importance biologique de ces petits ORFs dans la transposition des hélitrons

Le nombre d'occurrences dans les quatre différents clusters (Figure 3.10) est corrélé avec le nombre d'hélitrons autonomes. Plus un cluster possède d'hélitrons autonomes, plus le cluster a d'occurrences (Figure 3.10 et 3.12). Par exemple, le premier et le dernier cluster de la matrice d'occurrences sont les seuls clusters qui possèdent un grand nombre d'occurrences (plus de 100). Les hélitrons autonomes découverts (avec GENSCAN et BLASTP) correspondent aux hélitrons autonomes décrits par Kapitonov et Jurka [98]. Le troisième cluster ne contient qu'un seul hélitron autonome, mais ce dernier est commun avec le premier cluster. Le faible nombre d'occurrences du troisième cluster suggère que

la combinaison de termini de ce cluster s'est insérée dans l'hélitron autonome. L'hélitron autonome ne semble pas ou peu reconnaître la combinaison de ce troisième cluster. Le second cluster possède un seul hélitron autonome qui n'est pas partagé avec les autres clusters (Figure 3.12).

3.3.4 Une nouvelle nomenclature des hélitrons chez *Arabidopsis thaliana*

La Figure 3.10 montre qu'il y a 1369 combinaisons de termini qui peuvent représenter 1369 familles potentielles d'hélitrons dans une classification basée sur les extrémités. Cependant, les résultats précédents ont montré qu'une même séquence hélitronique pouvait correspondre à plusieurs combinaisons de termini insérées les unes dans les autres (Figure 3.13).

Le problème est alors de distinguer parmi tous ces termini si certaines combinaisons suffisent à expliquer l'ensemble des occurrences observées. Autrement dit, existe-t-il un ensemble restreint de combinaisons de termini qui soit caractéristique de l'ensemble des observations et qui ait pu servir de base à l'invasion du génome ?

D'un point de vue formel, il s'agit d'un problème d'optimisation combinatoire où l'objectif est de minimiser le nombre de sous-ensembles (séquences couvertes par un couple donné de termini) permettant de couvrir un ensemble O (l'ensemble des séquences d'hélitrons). Ce problème de couverture optimale est connu pour être un problème NP-difficile et il n'existe pas d'algorithme d'approximation polynômiale pour le résoudre [38].

Le jeu d'occurrences O a d'abord été défini en relation avec des hélitrons de taille maximale : une occurrence dans O débute avec une extrémité 5' de l'ensemble $LEFT$ et se finit avec une séquence 3' de l'ensemble $RIGHT$ et n'est pas incluse dans une autre occurrence. Les termini présents dans chaque occurrence sont alors examinés : chaque élément de O est associé avec l'ensemble des paires provenant de $C = LEFT \times RIGHT$ inclus dans cet élément. Comme il s'agit d'un problème NP-difficile, une heuristique doit être adoptée pour approximer le problème de couverture associé. L'algorithme glouton standard [38] lit les éléments de C et, à chaque étape, par une notion de maximum local du point de vue de la couverture, choisit une paire d'extrémités $CMax$. Cette paire couvre le plus grand nombre d'occurrences restant à chaque étape (section 2.5.2.

Nous avons cherché un compromis entre l'algorithme glouton et l'algorithme exhaustif qui explore toutes les solutions possibles. Ici les choix de $CMax$ sont dirigés comme dans l'algorithme glouton. Cependant, au lieu de retourner un seul choix ($CMax$), on en retient deux. Les deux alternatives que nous avons proposé à chaque étape sont :

- soit on prend $CMax$ ou la meilleure couverture des occurrences restantes (que nous nommons "double glouton").
- soit on prend $CMax$ ou la meilleure couverture qui remplace $CMax$ (que nous nommons "couverture CMax").

Pour choisir le meilleur algorithme, nous les avons implémentés et testés sur deux jeux de données : un jeu de données minimal qui correspond aux extrémités du chromosome 1 et le jeu de données contenant toutes les occurrences des hélitrons dans le génome entier d'*Arabidopsis thaliana*. Nous avons lancé 10 fois les trois algorithmes sur chaque jeu de données, et avons obtenu une valeur moyenne du temps de calcul. Le résultat de ce comparatif se trouve dans la Figure 3.14.

Nous avons observé que l'algorithme glouton, donne toujours le meilleur temps de calcul, mais donne toujours une solution moins bonne que les deux autres algorithmes. En comparant nos deux algorithmes, nous observons que l'algorithme "Couverture CMax" trouve une meilleure solution (ou au moins aussi optimale) que les deux autres. De plus, pour l'intégralité du génome, il obtient une meilleur solution en un meilleur temps de calcul que l'algorithme "double glouton". Nous avons donc choisi de conserver "Couverture CMax" que nous nommerons simplement SetCover dans la suite de cette thèse.

	Algo Glouton	Algo Cmax ou Autre Cmax	Algo Cmax ou meilleur couverture Cmax	
Temps CPU (sec)	15,4	15,63	15,71	Chromosome 1
Nombre de paires	30	29	29	Arabidopsis
Temps CPU (sec)	65,15	1942,15	426,68	Génome Entier
Nombre de paires	70	70	69	Arabidopsis

FIG. 3.14 – Comparatif de notre algorithme SetCover avec l'algorithme glouton et l'algorithme "double glouton". Nous avons comparé les trois algorithmes pour deux jeux de données : les couples d'extrémités hélitroniques du chromosome 1 et les couples du génome entier. Pour les deux jeux, nous avons comparé le temps CPU et le résultat de l'optimisation.

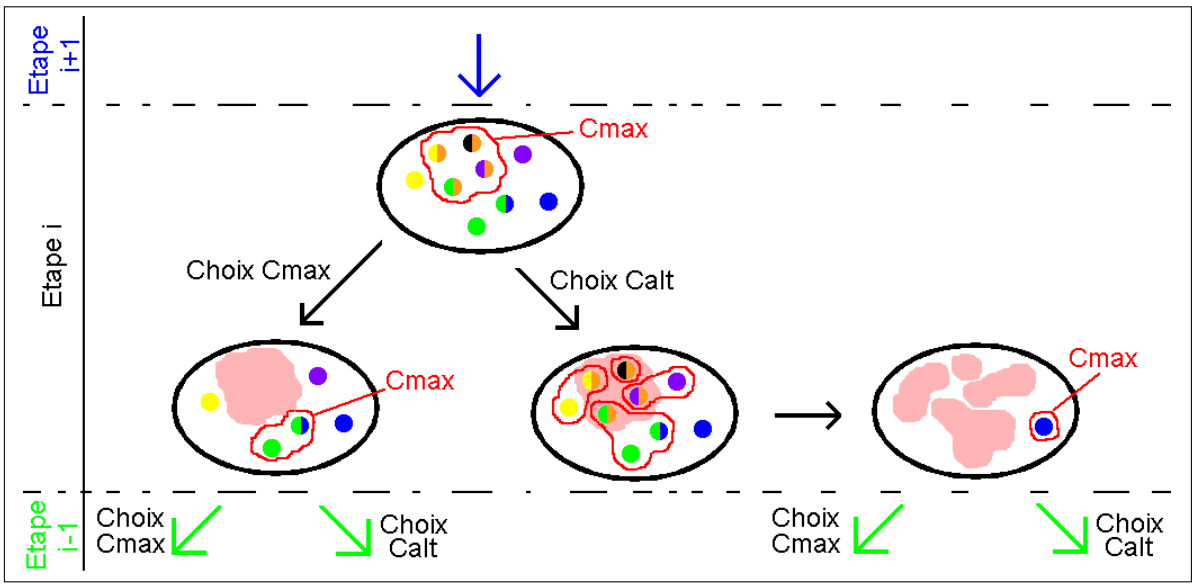


FIG. 3.15 – Fonctionnement de l'algorithme SetCover au cours de l'étape i. Les occurrences sont représentées par des points, et les couples sont représentés par les couleurs sur les points. A l'étape i, on repère le couple qui couvre le plus d'occurrences (cercle rouge). Soit on choisit ce couple CMax soit on choisit C_{Alt} : un ensemble de couples qui couvre les occurrences du couple CMax choisi précédemment (aire rouge clair). Ensuite, on élimine les occurrences couvertes par ces couples et on passe à l'étape $i+1$.

A chaque étape i (Figure 3.15), l'algorithme parcourt l'ensemble des occurrences O

et l'ensemble des couples C rattachés à ces occurrences pour trouver les séquences "singletons" (séquences n'ayant qu'un seul couple d'extrémité). Puis l'algorithme parcourt le reste des occurrences O et des couples C, pour récupérer le couple $CMax$, le couple ayant la plus grande couverture de O. L'algorithme parcourt une nouvelle fois les occurrences O restantes pour trouver les couples qui recouvrent la couverture de $Cmax$. Nous pouvons en déduire que l'algorithme a une complexité de $O(O^2.C.\log(O^2.C))$. L'algorithme détaillé est présenté ci-dessous :

Algorithm 2 Algorithme d'optimisation de Couverture SetCover({O},{C})

Require: Objets := $\{O_i\}$:= Ensembles des occurrences
Require: Couvertures C := $\{C_i\}$:= Ensemble des couples
Require: Match(O, C_j) := Sous-ensemble d'occurrences O contenant le couple C_j
Require: Termini(O_i, C) := Sous-ensemble de couples C contenu dans l'occurrence O_i
Ensure: MinCover := SetCover({O},{C})

MinCover := $\varnothing$
$Couvertures_1$:= Couvertures C
$Objets_1$:= Objets O
for i := 1 à la taille de {Objets O} **do**
  if taille(Termini(O_i, C)) = 1 (un singleton) **then**
    MinCover := MinCover + (Termini(O_i, C)
    $Couvertures_1$:= $Couvertures_1$ - (Termini(O_i, C)
    $Objets_1$:= $Objets_1$ - O_i
  end if
end for
C_{Max} := $ArgMax_{C_j \in Objets_1}$ (taille(Match$(Objets_1, C_j)$))
$Couvertures_2$:= $Couvertures_1$ - C_{Max}
$Objets_{Max}$:= Match$(Objets_1, C_{Max})$
$Objets_2$:= $Objets_1$ - $Objets_{Max}$

C_{Greedy} := **SetCover**($Objets_2$, $Couvertures_2$)
C_{AltMax} := **SetCover**($Objets_{Max}$, $Couvertures_2$)
$Couvertures_{AltMax}$:= $Couvertures_1$ - C_{AltMax}
$Objets_{AltMax}$:= $Objets_1$ - $\cup_{C_AltMax \in Couvertures_{AltMax}} Match(Objets_1, C_{AltMax}$

$C_{Alternative}$:= **SetCover**($Objets_{AltMax}$, $Couvertures_{AltMax}$)
if taille$(C_{AltMax} \cup C_{Alternative})$ < taille$(C_{Greedy} + 1)$ **then**
  return MinCover $\cup C_{AltMax} \cup C_{Alternative}$
else
  return MinCover $\cup C_{Max} \cup C_{Greedy}$
end if

Avant d'exécuter le programme SetCover, nous avons décidé d'éliminer toutes les combinaisons hélitroniques ayant moins de cinq occurrences et n'ayant aucun hélitron autonome détecté pour cette combinaison. Une analyse basée sur ces paires de termini a montré que ces couples semblent correspondre à des extrémités qui ont dégénérées en accumulant des mutations. On peut supposer que les mutations de ces extrémités empêchent leur

reconnaissance par les protéines RPA-hélicase. Les différentes occurrences présentes dans le génome proviendraient de différents événements de transposition d'hélitrons (reconnus par les protéines) qui auraient mutés et se seraient fixés au génome. Il est probable que ces occurrences vont alors continuer à muter et finiront par disparaître de la détection par un analyseur syntaxique.

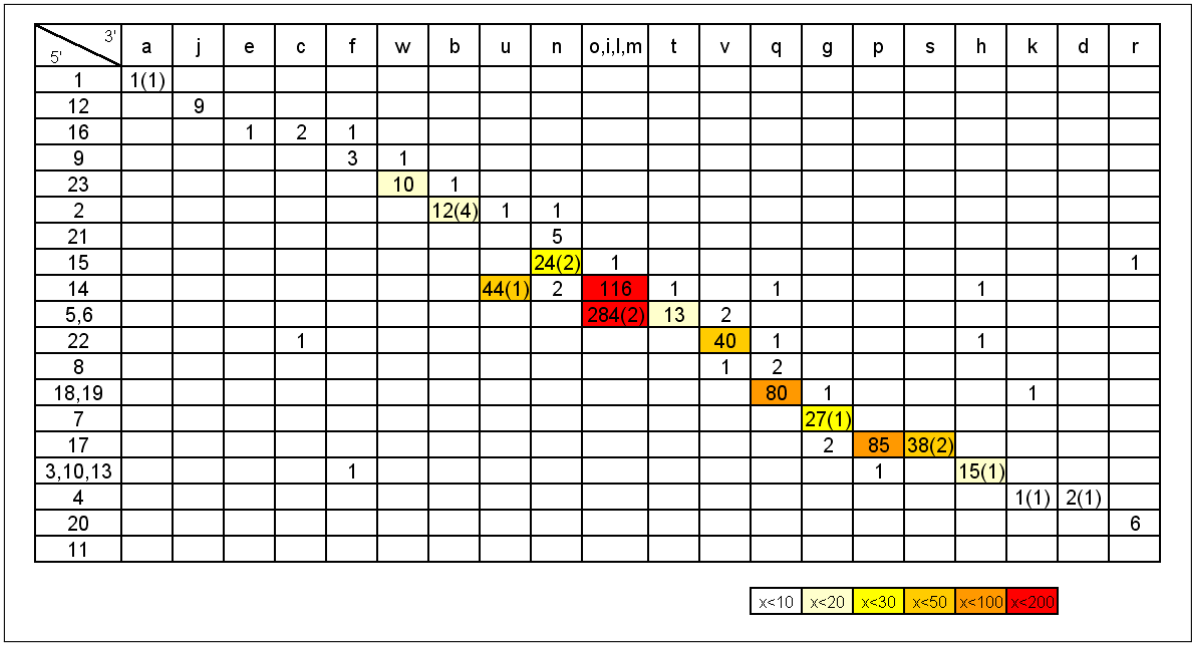


FIG. 3.16 – Matrice de combinaisons de termini qui recouvre l'ensemble des séquences hélitroniques chez *Arabidopsis thaliana*. Chaque cellule est colorée en fonction de sa fréquence. Chaque ligne représente un terminus 5' de 36 pb et chaque colonne un terminus 3' de 36 pb comme défini dans la figure 3.3.

Notre algorithme a retourné avec ces données filtrées 44 couples de termini qui ont recouvert l'ensemble des séquences dans le génome d'*Arabidopsis thaliana* (ensemble O) (Figure 3.16). Exceptés les couples 1_a et 12_j, tous les autres couples sont directement ou indirectement connectés à un hélitron autonome. La plupart des paires montrant un grand nombre d'occurrences ont au moins une extrémité en commun avec les hélitrons autonomes. Ces résultats sont en accord avec la nécessité pour un hélitron non-autonome donné d'utiliser les protéines de transposition codées par un hélitron autonome qui possède des extrémités similaires.

Nous avons proposé une nouvelle nomenclature à partir des 19 combinaisons de termini restantes (Figure 3.17). Les règles suivantes ont été choisies pour les familles : toutes les combinaisons qui contiennent des hélitrons autonomes sont appelées "Hélitron" suivi d'un nombre et toutes les autres combinaisons sont appelées "AtREP" suivi d'un nombre. Si les deux anciens noms d'extrémités [98] sont identiques et respectent la précédente condition, l'ancien nom de la famille est conservé.

Cette nomenclature montre beaucoup de nouvelles familles d'hélitrons autonomes (Hélitron6, 7, 8, 9, 10 dans la Figure 3.17). Néanmoins, un alignement multiple de ces nouveaux éléments autonomes avec les hélitrons autonomes de Repbase n'a pas montré de nouvelles séquences protéiques. Ces résultats montrent une dynamique et des mécanismes

Couple	Old name of 5' extremity	Old name of 3' extremity	New Name
1_a	Helitron1	Helitron1, Y1C	Helitron1
2_b	Helitron2, Y2	Helitron2, Y2	Helitron2
3,10,13_h	Helitron3, HelitronY1D, HelitronY3A	HelitronY1D	Helitron3
4_d	Helitron4	Helitron4	Helitron4
5,6_o,i,l,m	Helitron5, AtREP1, 2, 2A, 11, 11A, 14	HelitronY1E, AtREP1, 2, 2A, 3, 5, 11, 11A	Helitron5
7_g	HelitronY1A	HelitronY1B	Helitron6
4_k	Helitron4	HelitronY3A	Helitron7
17_s	AtREP6, 7, 8, 9, 13	AtREP13	Helitron8
14_u	AtREP3, 20	AtREP20	Helitron9
15_n	AtREP4	AtREP4, 15	Helitron10
12_j	HelitronY3	HelitronY3	AtREP1
14_o,i,l,m	AtREP3, 20	HelitronY1E, AtREP1, 2, 2A, 3, 5, 11, 11A	AtREP3
17_p	AtREP6, 7, 8, 9, 13	AtREP6, 7, 8, 9	AtREP6
18,19_q	AtREP10, 10A, 10B, 10C, 10D	AtREP10, 10A, 10B, 10C, 10D	AtREP10
20_r	AtREP12	AtREP12	AtREP12
5,6_t	Helitron5, AtREP1, 2, 2A, 11, 11A, 14	AtREP14	AtREP14
21_n	AtREP15	AtREP4, 15	AtREP15
22_v	AtREP21	AtREP21	AtREP21
23_w	AtREPX1	AtREPX1	AtREPX1

FIG. 3.17 – Nouvelle nomenclature des hélitrons. La première colonne correspond aux couples sélectionnés par l'optimisation. Les seconde et troisième colonnes correspondent aux anciens noms des extrémités des hélitrons [98]. La dernière colonne est le nouveau nom que nous proposons pour la famille d'hélitron.

combinatoires des extrémités pouvant générer de nouvelles familles d'hélitrons autonomes (voir section 3.3.5).

Nous avons voulu vérifier si les nouvelles familles chimériques, présentes dans notre nomenclature, étaient de "véritables" familles hélitroniques, c'est-à-dire capable de transposer. Les hélitrons s'insèrent dans d'autres éléments transposables ou hélitrons [98]. Il est donc possible de prouver *in silico* qu'un élément transposable est capable de transposer, si l'on détecte un élément transposable i inséré dans un élément répété j et que ce dernier est présent à un autre locus sans la séquence i. L'élément j sans i est appelé une "niche" vide [98]. Nous avons recherché ces niches vides pour ces nouvelles familles d'hélitrons. Nous avons récupérer les 50 nucléotides positionnés en 5' et les 50 nucléotides positionnés en 3' pour chaque séquence hélitronique et avons concaténé ces deux bordures. Pour chaque séquence, nous avons exécuté CENSOR sur le site Repbase (www.girinst.org/censor/index.php) [107]. Nous avons considéré qu'un alignement créé par CENSOR était significatif s'il mesurait plus de 75 pb et avait une similarité de séquence supérieure à 75 %.

La Figure 3.18 montre un cas de site d'insertion pour chaque nouvelle famille hélitroniques et les niches vides correspondantes. Ce résultat confirme que les nouvelles familles découvertes sont capables de transposer.

3.3.4.1 Relation entre les familles et les hélitrons autonomes

Si les protéines de transposition RPA-hélicase reconnaissaient un motif hélitronique non spécifique, il n'y aurait pas de corrélation entre le nombre d'autonomes d'une famille donnée et l'amplification de cette famille. Alors que la comparaison des séquences internes n'a donné aucune corrélation entre les hélitrons autonomes et non-autonomes [98]. Notre

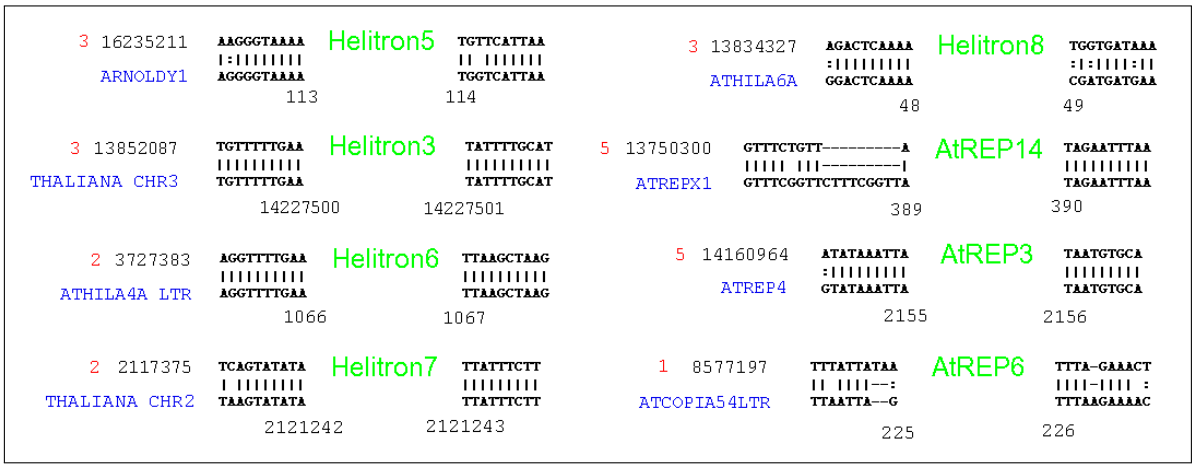


FIG. 3.18 – Comparaison de séquences flanquantes des nouvelles familles hélitroniques et de niches vides. La nom des niches est indiqué en bleu. Les noms des familles hélitroniques sont colorés en vert. La position de la séquence flanquante est indiquée avant la séquence ; le numéro de chromosome est indiqué en rouge.

analyse basée sur les termini a mis en évidence des relations significatives entre certains hélitrons autonomes et non-autonomes qui pouvaient être classés dans la même famille (Figure 3.17).

De plus, la corrélation observée entre la présence d'autonomes et l'amplification des non-autonomes possédant les mêmes extrémités suggère fortement que les protéines RPA-hélicase reconnaissent préférentiellement les termini d'hélitrons qui sont similaires à ceux de l'autonome. Cependant, un certain nombre de familles, telle que la nouvelle famille AtREP3, est exclusivement constituée d'éléments non-autonomes. Mais à l'exception de la famille AtREPX1, toutes ces familles ont au moins une extrémité (5' ou 3') en commun avec une famille autonome. Par exemple, la famille AtREP3 (combinaison 5' 14_o,i,l,m 3') partage l'extrémité 3' avec la nouvelle famille autonome Hélitron5 (5' 5,6_o,i,l,m 3').

De précédentes études sur le transposon bactérien IS91, qui utilise aussi le mécanisme de réplication rolling-circle et qui possède un ORF de type hélicase [16, 47], ont montré que seule l'extrémité contenant la tige-boucle subterminale est nécessaire et suffisante pour sa transposition [145]. Si le même mécanisme fonctionne aussi dans le génome d'*Arabidopsis thaliana*, la présence d'extrémités 3' communes peut expliquer l'amplification d'hélitrons non-autonomes telles que les familles AtREP3, AtREP10 et AtREP15 (Figure 3.17) par des hélitrons autonomes d'autres familles. L'amplification de ces familles d'hélitrons non-autonomes, sans autonome associé, peut aussi avoir été effectuée par des hélitrons autonomes ancestraux qui ne peuvent plus être détectés par l'identification d'ORF et l'analyse des séquences dans le génome actuel. L'absence de ces hélitrons autonomes peut aussi simplement être la conséquence de leur dégradation au cours de l'évolution

3.3.5 Caractérisation des hélitrons chimériques

Beaucoup d'occurrences d'hélitrons tronqués contenant une seule extrémité ont été observés dans le génome d'*Arabidopsis thaliana* (Figure 3.7 et 3.8), suggérant ainsi qu'ils sont le résultat d'une transposition incomplète. D'un autre côté, il y a aussi un nombre significatif d'hélitrons présentant des combinaisons de termini de deux ou de plusieurs familles distinctes pour une séquence unique (Figure 3.10 et 3.12). Finalement, le modèle de séquence hélitronique est bordé par une seule extrémité 5' et une seule extrémité 3' (Figure 3.3). Il est donc extrêmement difficile de proposer une classification homogène des hélitrons prenant en compte à la fois la variabilité de la séquence interne et la dynamique des termini 5' et 3'. Néanmoins, cette structure combinatoire d'hélitrons peut correspondre à d'importantes propriétés biologiques. Ainsi, l'insertion d'hélitrons tronqués dans le voisinage d'autres hélitrons (entiers, tronqués ou chimériques) peut être une source de variabilité structurale, attribuée au fonctionnement des protéines de transposition. Si la transposition des hélitrons opère bien par reconnaissance des extrémités terminales, il est naturel d'observer ce phénomène d'extrémité hélitronique tronquée avec une extrémité d'un hélitron entier ou d'un autre hélitron tronqué [145, 119] (Figure 3.19). Les résultats des Figures 3.10 et 3.12 suggèrent que l'utilisation de telles combinaisons de termini lors de la transposition est possible, bien que certaines combinaisons soient transposées préférentiellement (Figure 3.10).

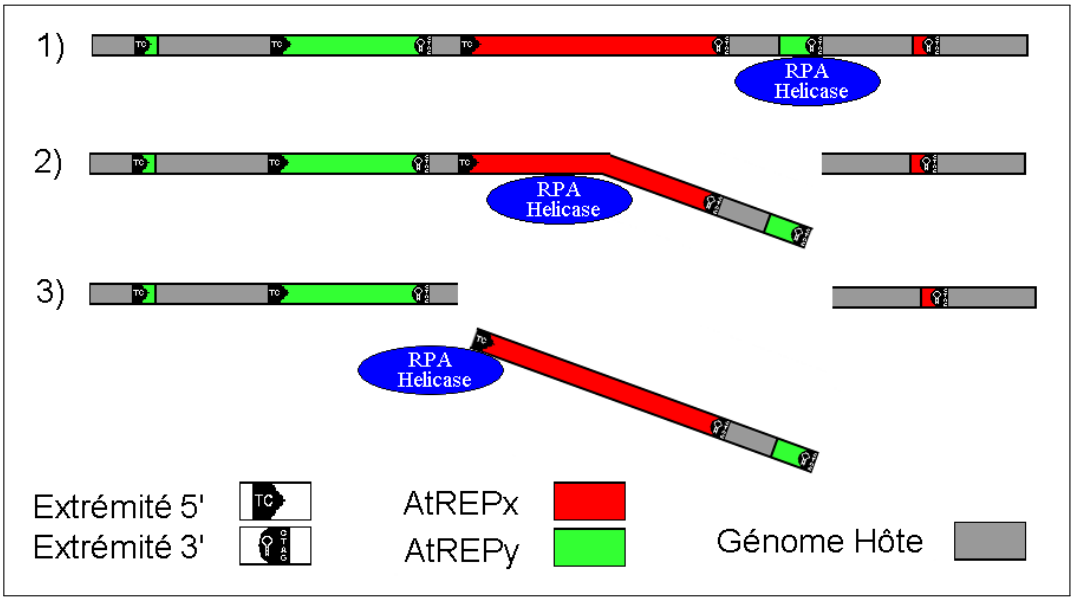


FIG. 3.19 – Hypothèse de mécanisme moléculaire de la création d'hélitron chimérique (adapté de Feschotte *et al.* [70] et Gutierrez et al [77]). (1) : Un hélitron entier (couleur rouge) est situé à proximité de plusieurs hélitrons tronqués. Les protéines de transposition [98, 70, 77] reconnaissent l'une des extrémités 3' d'un hélitron tronqué AtREPy. (2) : Les protéines coupent le terminus 3' de l'hélitron tronqué et se déplacent ensuite vers une extrémité 5'. (3) : Les protéines RPA-hélicase reconnaissent l'extrémité 5' de l'hélitron complet ATREPx. La séquence est coupée et l'hélitron chimérique est transposé.

Les hélitrons tronqués peuvent être un important mécanisme de modularité de la séquence interne des hélitrons et/ou de la création de chimères (Figure 3.19). De plus, comme le montre la Figure 3.19 et comme observé chez le maïs [149], cette variabilité et

ces combinaisons de séquences peuvent récupérer des fragments d'ADN génomique qui seront mobilisés avec la transposition de l'hélitron. Par exemple l'ORF "LARKLPVTQ-KEYSKTQTLI" non identifié par BLASTP est présent dans de nombreux hélitrons et aussi présent à la position 18749092 à 18749149 du chromosome 1 où nous n'avons pas détecté de séquences hélitroniques.

Résumé

La modélisation syntaxique de l'hélitron basé seulement sur la nature des termini a donc permis l'identification et la classification de tous les hélitrons présents chez *Arabidopsis thaliana*. Cette identification a aussi permis la découverte de nouvelles familles, principalement par recombinaison de termini. L'identification exhaustive des hélitrons a montré que la majorité des séquences hélitroniques est composée par de multiples combinaisons de termini, suggérant qu'une seule séquence hélitronique pouvait être déplacée par plusieurs protéines de transposition différentes. Les résultats précédents ont enfin permis de comprendre la relation entre les hélitrons autonomes et non-autonomes de ce génome (Figure 1.24).

Contribution scientifique

Article : Tempel S. Nicolas J. El Amrani A. Couée I. Model-based Identification of Helitrons Results in a New Classification of Their Families in *Arabidopsis thaliana*. GENE (accepté après révisions).

Conférence : Tempel S., Nicolas J., Couée I, and El Amrani A. 2005. 1st International Conference/Workshop : Genomic Impact of Eukaryotic Transposable Elements. Combinatorics of helitron termini in *Arabidopsis thaliana* genome reveals strongly structured superfamilies. Asilomar Californie.

Conférence : Tempel S., Nicolas J., Couée I, and El Amrani A. 2006. XIV ème Colloque Eléments Transposables. Combinatorics of helitron termini in Arabidopsis thaliana genome reveals strongly structured superfamilies. Clermont-Ferrand.

Poster : Tempel S, Couée I, El Amrani A and Nicolas J. 2006. Etude de l'évolution d'éléments génétiques mobiles : application à la découverte de superfamilles d'hélitrons chez A. thaliana. JOBIM. Bordeaux.

3.4 DomainOrganizer : identification de la modularité des éléments transposables

Au début de la thèse, nous avions réalisé un alignement multiple de l'ensemble des copies de la famille AtREP3 (Figure 3.20). Cet alignement a montré que la diversité des hélitrons n'était pas seulement due à des mutations ponctuelles, mais était surtout due à des mutations par blocs de séquences.

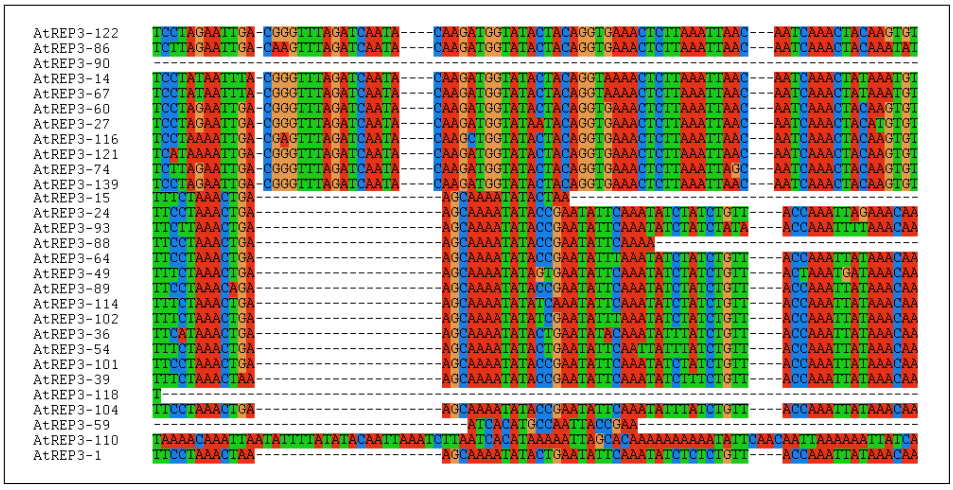


FIG. 3.20 – Zone d'un alignement réalisé par ClustalW [200] de la famille AtREP3. Les nucléotides sont colorés par ClustalW.

Cette présence de structure en blocs des hélitrons n'est pas un cas isolé, beaucoup d'études d'éléments transposables évoquent une structure en blocs, modules ou domaines [126, 202]. Par exemple, les éléments LINEs et SINEs sont décrits comme des rétroéléments composés d'un domaine A et B [49]. Hormis pour quelques familles qui dérivent d'ARNt [109], l'insertion/délétion des blocs observée dans l'alignement multiple brouille toutes les études évolutives basées sur la phylogénie. Ainsi, aucune étude ne caractérise exhaustivement l'évolution des éléments transposables non-autonomes par substitutions/insertions/délétions.

Nous avons voulu savoir si les blocs de domaines découverts dans l'alignement multiple étaient dus à des phénomènes aléatoires ou à des phénomènes biologiques. Nous avons alors étudié et caractérisé manuellement l'ensemble des blocs présents dans la famille AtREP3 (Figure 3.21). Cette étude a permis de montrer que les séquences internes des copies de la famille AtREP3 n'étaient pas un assemblage aléatoire de séquences nucléiques. Une première analyse manuelle de ces domaines présents dans la séquence interne des AtREP3 nous a permis d'expliquer globalement l'histoire évolutive de ces hélitrons. Néanmoins, la caractérisation manuelle de ces domaines était par essence fastidieuse et difficilement reproductible sur d'autres familles. Nous avons donc travaillé sur une méthode automatisée de caractérisation et de visualisation de ces domaines internes dans n'importe quelle famille d'éléments transposables.

Il existe des méthodes qui détectent des fragments conservés dans les séquences génomiques telles que MAUVE [43] ou REPUTER [116]. Ces méthodes sont plutôt adaptées à l'étude de grandes segmentations de nucléotides, et ne sont pas prévues pour une analyse

détaillée de blocs d'ADN contenant beaucoup d'erreurs. A l'exception de MOSAIC [6], aucun outil n'était adapté à une détection fine de l'architecture des séquences répétées. Malheureusement, l'utilisation de MOSAIC ne convient pas à la recherche des blocs dans la séquence interne des hélitrons, car il ne détecte pas les répétitions de domaines internes et ne segmente pas les séquences en domaines mais permet seulement une identification des domaines homologues.

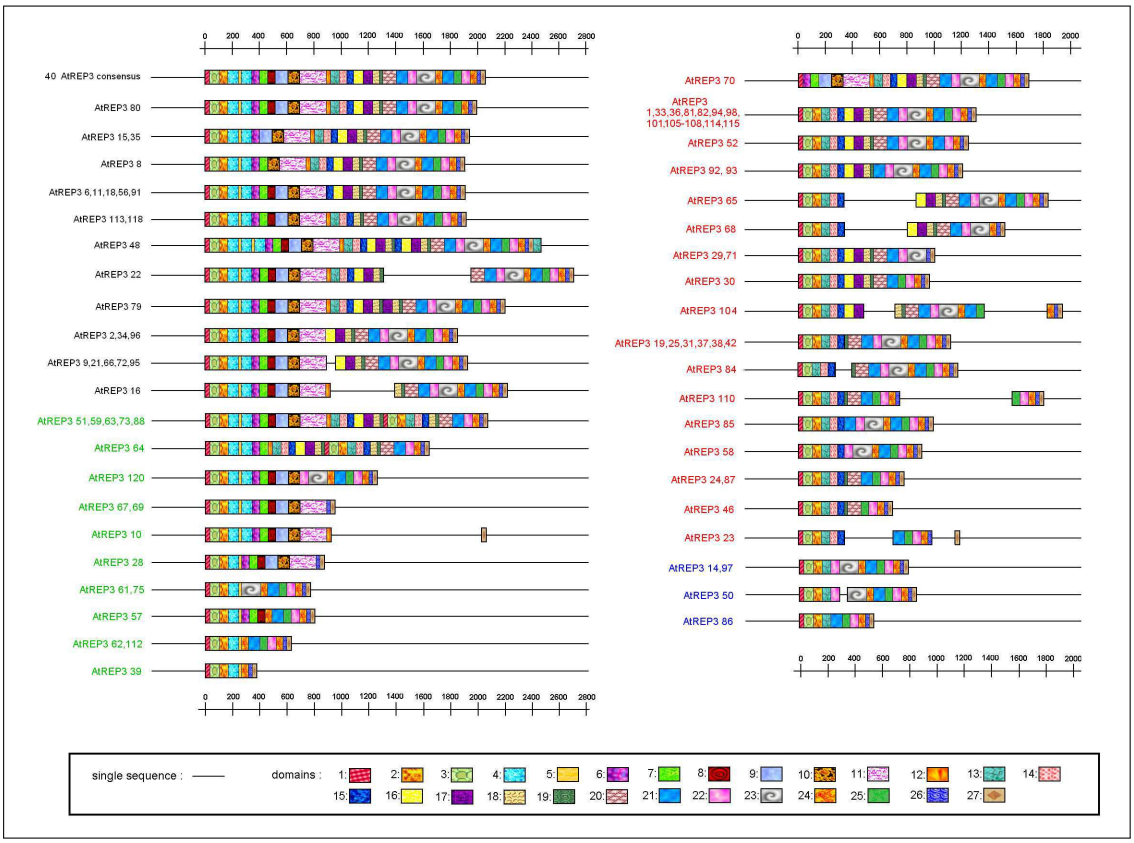


FIG. 3.21 – Visualisation de l'organisation en domaines des éléments AtREP3. Chaque domaine répété possède une texture unique. Les insertions de séquence uniques sont représentés par une simple ligne. Les différentes couleurs des noms des AtREP3 correspondent à une classification manuelle des AtREP3 en sous-groupes.

Pour étudier cette dynamique par blocs des séquences internes des hélitrons, nous avons proposé une méthode et avons créé un outil nommé DomainOrganizer [196] qui a donc été conçu pour détecter l'organisation en domaines des éléments répétés.

3.4.1 Présentation générale de la méthode de détection des domaines nucléiques

Par analogie avec les protéines [185], nous appelons "domaine nucléique" un fragment d'ADN répété dans plusieurs séquences. Les domaines nucléiques peuvent avoir des fonctions codantes (e.g. les exons), mais aussi des fonctions structurales telles que les TIR palindromiques des éléments transposables de classe II.

DomainOrganizer représente une famille d'éléments répétés comme une combinaison de domaines nucléiques. Pour obtenir ces combinaisons, la méthode associe des algorithmes existants et des algorithmes originaux (Figure 3.22). D'abord, l'ensemble des séquences est aligné avec ClustalW [200] ou Dialign [150]. Un algorithme original (DomainDetection) (Figure 3.22) détecte l'ensemble des frontières possibles entre domaines dans l'alignement. HMMER [59] produit une caractérisation, de chaque domaine potentiel, basée sur des profils HMM (Hidden Markov Model). Un nouvel algorithme original (DomainOptimizer) extrait le sous-ensemble de domaines produisant un recouvrement optimal des séquences. CHAVL [128] classe ensuite les séquences par rapport à la présence/absence des domaines et finalement un outil original (DomainRender) crée un graphe de toutes les séquences segmentées en domaines avec leur arbre de classification (Figure 3.22).

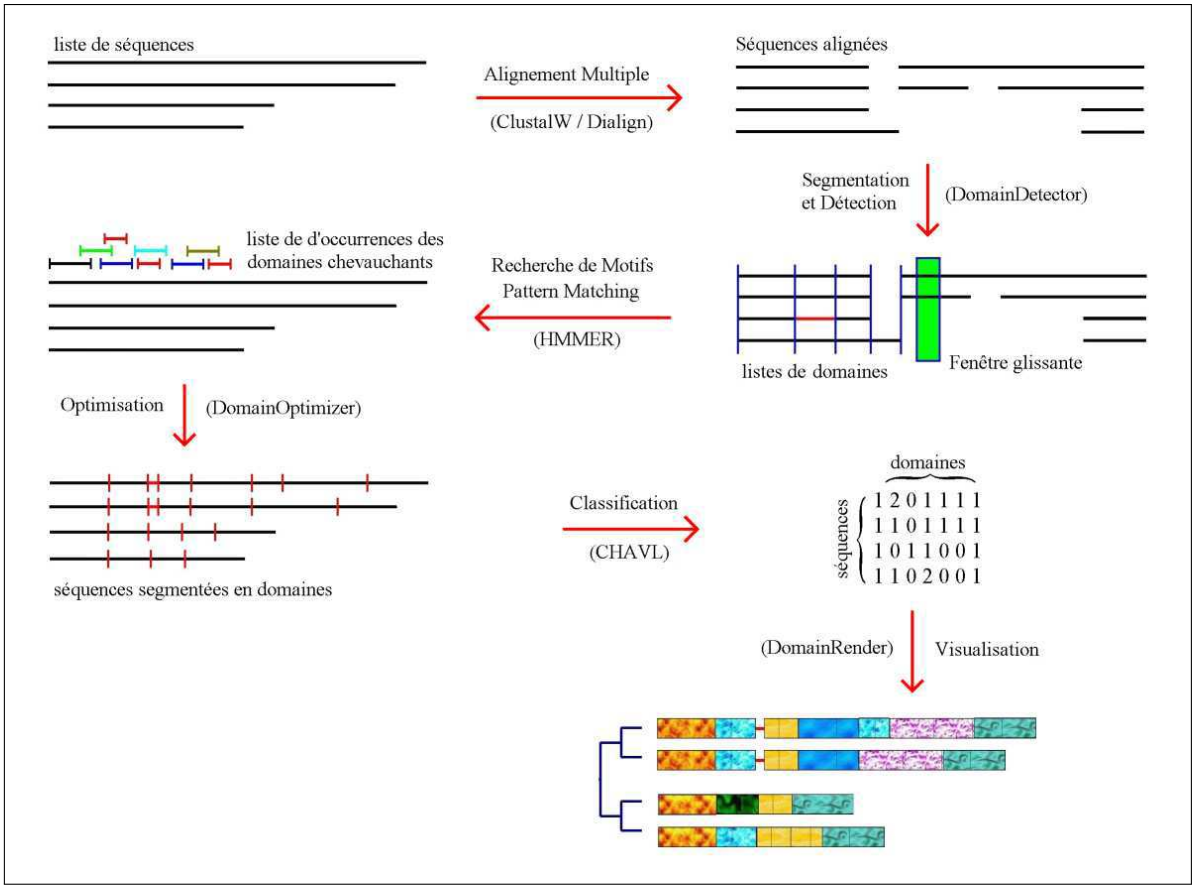


FIG. 3.22 – Principales étapes de détection et de visualisation des domaines présents dans un ensemble de séquences répétées. Les programmes utilisés sont indiqués entre parenthèses. Après l'alignement des séquences, DomainDetector trouve tous les domaines potentiels. HMMbuild et HMMcalibrate [59] créent les profils HMM de chaque domaine et Hmmsearch localise ces domaines dans toutes les séquences. DomainOptimizer extrait le minimum de domaines couvrant toutes les séquences. Cette détection crée une matrice de présence/absence de domaines, que CHAVL [128] utilise pour classer les séquences. Enfin, DomainRender crée l'image de la famille d'éléments incluant la classification et la localisation des domaines.

La question de la segmentation des grandes séquences d'ADN est étudiée depuis que les séquences génomiques sont devenues disponibles. Il y a un intérêt particulier pour la

détection d'isochores, d'îlots CpG, de signaux d'origine ou de terminaison de la réplication ou encore dans la détection des régions codantes/non codantes [131]. Toutes ces méthodes basées sur l'analyse informatique de sous-séquences utilisent des statistiques. Les deux principales méthodes sont l'estimation de paramètres caractéristiques de chaque segment et la segmentation par HMM. La méthode d'estimation utilise des matrices de poids, où chaque matrice correspond à un état différent de la séquence : par exemple, les séquences codantes et les introns [131]. Elle parcourt la séquence qui est vérifiée à chaque position avec les différentes matrices. La similarité entre une matrice de poids et une position donnée permet de connaître l'état de la séquence à cette position [179]. La segmentation marque les blocs avec les meilleurs modèles le long de la séquence. La segmentation binaire récursive trouve la meilleure coupe d'une séquence en deux sous-séquences avec le calcul de deux indices, l'un pour mesurer le choix de la coupe (divergence de Jensen-Shannon) et l'autre pour stopper la récursivité [9, 17, 131, 157].

Toutes ces études ont cherché une décomposition "brute" des séquences, basée sur leur contenu statistique. Néanmoins, il est possible de considérer le problème de segmentation à un niveau plus fin où les segments sont caractérisés par un ensemble de mots (sous-séquences) qui représente un langage. L'analyse de l'architecture des domaines d'une séquence peut être énoncée comme un problème d'optimisation. Gionis et Mannila ont appelé "problème $(k, h) - segmentation$", le problème qui consiste à trouver la meilleure segmentation d'une séquence de longueur n en k segments, chaque segment appartenant à un ensemble de sources h avec $h \leq k$ [73]. Ce problème est NP-complet dans des conditions assez larges.

Nous proposons un problème de segmentation légèrement différent, fondé sur l'hypothèse que les sources (les transposons) sont des séquences qui ont été copiées et qui ont divergées dans le génome. En effet, les copies d'une famille d'éléments transposables ont un ancêtre commun et leurs différences proviennent des mutations subies après leur copie. A partir d'une famille de copies de transposons S, nous avons recherché l'ensemble minimum de sous-séquences $D = \{Di\}$ (les domaines) tel que S puisse être exprimé comme une concaténation d'éléments de D. Ce minimum représente à la fois le nombre minimal de domaines D différents pour l'ensemble des séquences mais aussi le nombre minimal de domaines utilisé par séquence. Par exemple, soit une séquence S_1 composée par ATTTGA et une séquence S_2 composée par AATGA. Le minimum de sous-séquences qui peut représenter S_1 et S_2 est $D_1 = A$, $D_2 = T$ et $D_3 = TGA$. Nous pouvons représenter les deux séquences par : $S_1 = D_1.D_2.D_2.D_3$ et $S_2 = D_1.D_1.D_3$. Les trois nucléotides A, G et T ne sont pas choisis comme sous-séquences minimales, car si leur nombre global de domaines est identique, le nombre de domaines par séquence est supérieur. Puisque la segmentation prend en compte l'ordre de la séquence et les autres copies de la séquence en parallèle, nous nous attendons à une division plus fine que celle obtenue par l'analyse d'une séquence unique.

Nous limitons la taille inférieure des domaines à $m = MinSizeDomain$ qui est la taille minimale et suffisante pour caractériser de façon non ambiguë chaque domaine. Plus précisément, chaque domaine est caractérisé par un mot w de taille supérieure à m. Chaque occurrence du domaine contient un mot qui peut différer de w par au maximum $MaxErrors$ erreurs. Une erreur est une substitution, une délétion ou une insertion à une position donnée dans un alignement multiple. Il est possible qu'aucune occurrence ne corresponde exactement à w. Par exemple, soit deux séquences $S_1 = GGAATGACCTGA$ et $S_2 = GGATTAACCTGA$, $MinSizeDomain = 5$ et $MaxErrors = 1$, la sous-séquence

AATGA contenue dans S_1 et ATTAA contenue dans S_2 sont deux occurrences du même domaine $w = ATTGA$ avec une erreur de substitution (Figure 3.23).

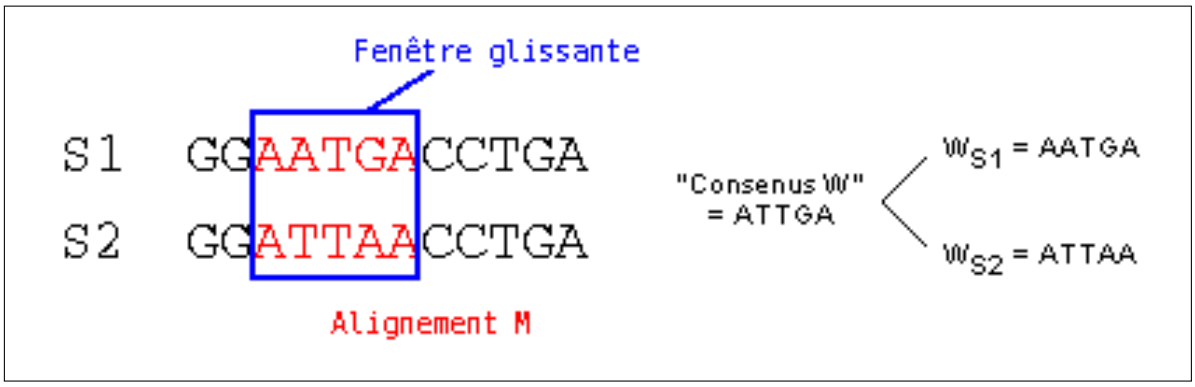


FIG. 3.23 – Exemple de domaine w dans des séquences S_1 et S_2. Le cadre bleu représente la fenêtre d'alignement de taille $MinSizeDomain = 5$. Les lettres en rouge représentent l'alignement M contenu dans cette fenêtre.

Notre algorithme est d'abord basé sur l'alignement multiple des séquences et sur la détection d'extrema locaux par une fonction d'hétérogénéité. Cette fonction est construite sur deux indices : $NbEmpty(M)$ et $NbDiff(M)$. Elle mesure si, à une position donnée dans un alignement M, le nombre de domaines contenu dans la fenêtre glissante change.

Soit L la longueur de l'alignement multiple des séquences, cette longueur est généralement plus grande que la plus grande des séquences qui compose l'alignement. L'algorithme lit une portion de l'alignement de taille $MinSizeDomain$ le long d'une fenêtre glissante sur les séquences alignées. Cette fenêtre lit toutes les séquences de la position i à $i + MinSizeDomain$ tel que $1 \leq i \leq L - MinSizeDomain$ (Figure 3.23). Nous avons appelé alignement M la portion de l'alignement lue par le programme. $NbEmpty(M)$ est défini comme le nombre de séquences dans M représenté par un gap (le nombre de séquences avec $MinSizeDomain$ "-" caractères dans M).

$NbDiff(M)$ est défini comme le nombre maximum de positions qui diffère du $consensus(M)$ observé pour les mots en M. Par exemple, pour la figure 3.23, $NbDiff(M)$ vaut 1 : il n'y a qu'une erreur de substitution entre les mots lus par la fenêtre glissante et le mot consensus w obtenu à partir de M. En général, les variations observées correspondent à des mutations ponctuelles et il est possible qu'il n'existe qu'un seul domaine dans l'alignement M si le nombre de positions conservées est assez élevé pour toutes les séquences (moins de $MaxErrors$).

$consensus(M)$ est la notion standard d'un mot consensus excluant les gaps : la $i^{ème}$ lettre du mot $consensus(M)$ sera la lettre la plus fréquente dans la liste des lettres (moins la lettre "-") à la position i dans M. Par exemple, soient quatre séquences alignées, la fenêtre lit l'alignement M : $S_1 = ATTGA$, $S_2 = ATT-A$, $S_3 = AATGA$ et $S_4 = ATTAA$. A la quatrième position de M, l'algorithme lit deux G (S_1 et S_3), un A (S_4) et un "-". La quatrième lettre du consensus sera la lettre G. De cette façon, de faux domaines sont évités et chaque domaine est caractérisé par un mot au moins de la taille de m. Un alignement multiple avec un coût faible d'insertion et d'extension de gap favorise la création de gaps quand les séquences ne sont pas similaires. Cette création limite le nombre de séquences

divergentes à l'intérieur de m et favorise la création d'un unique $consensus(M)$ dans M. On doit noter que $NbDiff$ et $MaxErrors$ sont liés à la taille de m et que, pour un alignement M où au moins l'une des séquences n'est composée que de "-", $NbDiff$ vaut m.

La fonction hétérogénéité est un codage dans la base $m+1$ des deux indices, avec une priorité accordée sur l'indice $NbEmpty$. Pour un alignement M de la longueur m, nous obtenons :

$$Heterogeneity(M) = (m+1) \cdot NbEmpty(M) + NbDiff(M)$$

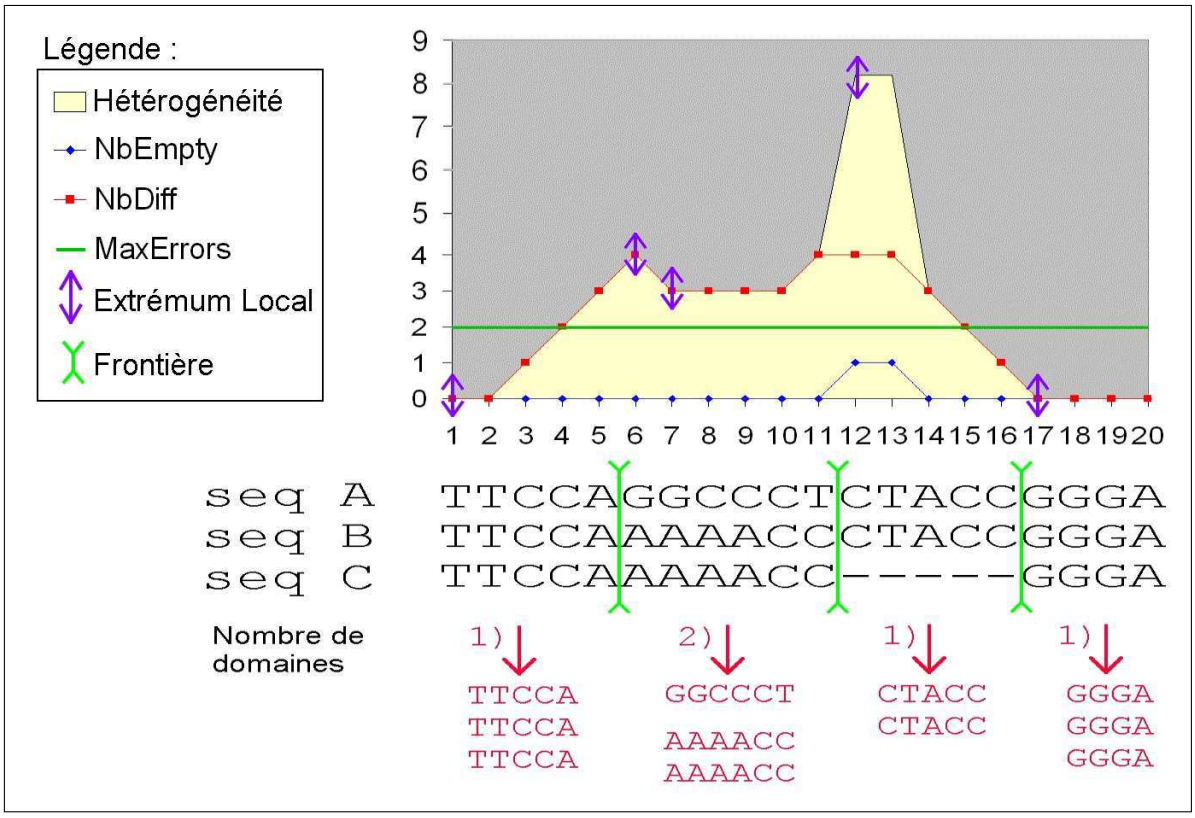


FIG. 3.24 – Exemple de détection de domaines. $MinSizeDomain$ vaut 4 et $MaxErrors$ vaut 2. Une fenêtre glissante de taille $MinSizeDomain$ parcourt l'alignement calculant à chaque étape $NbDiff$ et $NbEmpty$. $NbDiff$ vaut 0 à la position 1. Le nombre d'erreurs augmente jusqu'à un extremum local de 4 erreurs à la position 6. Le domaine détecté correspond à la position 1 à 5 (=6-1). La fenêtre repart à la position 9 (=5+4). La fenêtre glisse jusqu'à la position 12. Une séquence gap apparaît à cette position. Une nouvelle frontière est détectée avec deux domaines s'étendant de la position 6 à 11. A la position 14, $NbEmpty$ vaut 0 et $NbDiff$ vaut 2 et le nombre d'erreurs diminue jusqu'à la position 17. A cette position, $NbDiff$ et $NbEmpty$ valent 0, c'est un extremum. Le segment détecté va de la position 12 à 16. Le dernier segment de 17 à 20 est le dernier domaine.

Finalement, pour un ensemble donné de séquences, l'algorithme détecte les frontières des domaines comme les extrema locaux de la fonction d'hétérogénéité, c'est-à-dire quand $NbEmpty$ change, ou quand $NbDiff$ atteint $MaxErrors$ entre deux frontières. La fi-

gure 3.24 montre un exemple de détection de domaines et du calcul de la fonction "hétérogénéité" dans un alignement multiple. La fonction d'hétérogénéité est utilisée dans l'algorithme DomainDetector suivant :

Algorithm 3 Algorithme de DomainDetector

Require: Alignement, MinSizeDomain et MaxErrors
  SeqDomaines = ∅
  début := 0
  for i := 1 à la Longueur_des_Séquences - MinSizeDomain **do**
    Calcule(hétérogénéité(Alignement[i : i+MinSizeDomain]))
    if la valeur hétérogénéité a atteint un extremum local **et** NbEmpty a changé
    ou le signe |NbDiff-Maxerrors| a changé depuis le précédent extremum **then**
      SeqDomains := SeqDomains ∪ ([début, i+MinSizeDomain]
      ClasserSéquences(Alignement[début : i+MinSizeDomain], MaxErrors))
      début := i + MinSizeDomain
    end if
  end for
  return SeqDomains

  ClasserSéquences(M, MaxErrors) =
  if M = ∅ **then**
    return ∅
  else
    return (S' = u ∈ M / distance(u, consensus(M)) < MaxErrors) ∪ ClasserSéquences(M S', MaxErrors)
  end if

Après la détection des frontières dans l'alignement multiple, l'algorithme dispose d'une liste de domaines qui recouvrent l'ensemble des séquences S. Néanmoins, l'alignement multiple ne détecte pas les sous-séquences répétées à l'intérieur d'une même séquence et DomainDetector considère ces sous-séquences comme des domaines distincts. Par exemple, si un domaine est présent en deux copies dans la séquence A et en une seule copie dans la séquence B, alors l'alignement multiple n'alignera qu'un seul domaine de la séquence A avec la séquence B et DomainDetector considérera la deuxième copie du domaine dans la séquence A comme un nouveau domaine. De plus, chez les éléments transposables non-autonomes, il y a un fort remaniement de séquences internes entraînant des inversions de sous-séquences ou l'insertion de séquences complémentaires inverses [89]. Dans ce cas aussi, DomainDetector considère le domaine inversé ou complémentaire inverse comme un nouveau domaine.

Pour pallier ce problème, nous recherchons ensuite la séquence inverse et complémentaire inverse de chaque séquence. Chaque domaine "normal", sa séquence inversée et sa séquence complémentaire inversée sont transformés en profil HMM. Le profil HMM est créé par HMMbuilt (composant de HMMER) à partir de l'alignement multiple de toutes les occurrences d'un domaine [59]. Ensuite le profil est recherché dans toutes les séquences. Cette recherche a permis de retrouver les domaines répétés, inversés ou complémentaires inversés à l'intérieur d'une même séquence. Si deux domaines distincts (non inversés et non complémentaires-inversés) s'apparient avec HMMSearch à des positions identiques,

toutes les occurrences des deux domaines sont alignées et l'algorithme conserve la séquence consensus des occurrences des deux domaines. Si un domaine inversé ou complémentaires-inversé fusionne avec un autre domaine sur le brin sens, ils sont alors considérés comme identiques.

Même si l'étape des HMM a éliminé des domaines redondants, il reste encore plus de domaines que nécessaire pour recouvrir l'ensemble des séquences. Une étape d'optimisation de recouvrement des domaines est réalisé (Figure 3.22). Cette étape vise à réduire au minimum la distance entre les domaines successifs d'une séquence et réduire au minimum le nombre global de domaines utilisés dans l'ensemble des séquences. Cette minimisation d'un ensemble d'occurrences de domaines, sur un ensemble de séquences, est un problème clairement NP-complet.

L'ensemble des occurrences des domaines de D dans une séquence S_l de S forme un graphe orienté G_l, appelé *graphe de couverture*, si l'on décide de relier deux occurrences lorsqu'elles peuvent se suivre dans une segmentation valide (Figure 3.25). A ce graphe, on ajoute un sommet initial et un sommet terminal virtuels s_l et t_l et les arcs $s_l \rightarrow d$ ($d \rightarrow t_l$) pour tout domaine d pouvant débuter (terminer) une segmentation. G_l est un graphe acyclique et tout chemin de s_l à t_l représente une segmentation valide pour la séquence S_1. Les arcs de G_l sont pondérés : soit un arc $i \rightarrow j$ de G_l, nous notons (d_i, f_i) et (d_j, f_j) les positions de début et de fin des occurrences i et j. Le poids de $i \rightarrow j$ est $c_{ij}^l = |d_j - f_i|$ (Figure 3.25). Tous les sommets ne sont pas reliés à tous les autres sommets : pour relier deux sommets, il faut que le poids (ou distance) soit inférieur à la taille minimale d'un domaine ($c_{ij}^l \leq MinSizeDomain$). De plus, nous avons pondéré chaque sommet par la taille du domaine.

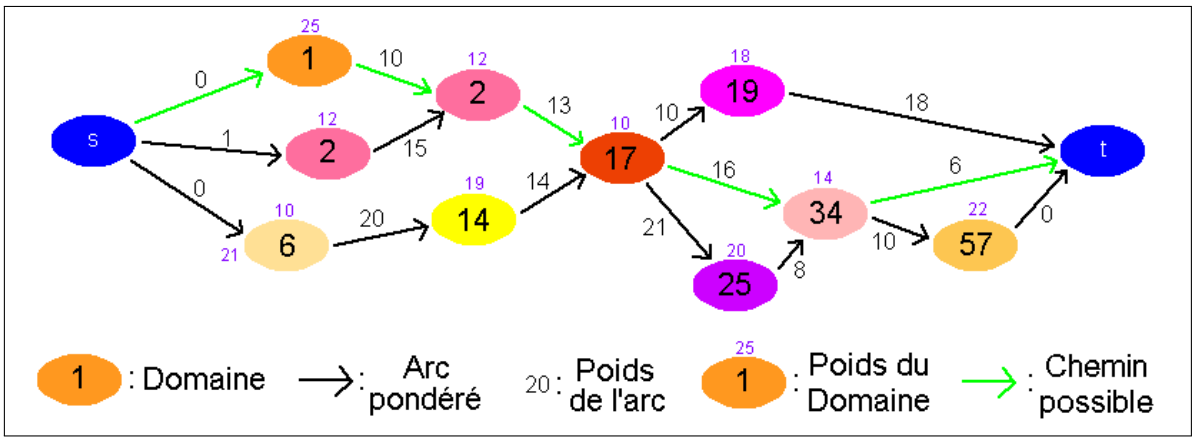


FIG. 3.25 – Exemple de graphe de couverture pour une séquence S_1. s et t sont les sommets virtuels qui commencent et finissent une segmentation. Le poids des arcs représente la distance, en valeur absolue, entre le début du sommet de l'arc entrant et la fin du sommet de l'arc sortant.

Le problème est de représenter le plus complètement possible les séquences de S à l'aide d'occurrences de domaines, tout en utilisant un minimum de domaines différents. Nous cherchons un chemin de s_l à t_l dans chaque graphe G_l en minimisant la somme des poids obtenus pour l'ensemble des séquences. De plus, si un domaine k est utilisé dans une séquence i, le poids du domaine sera nul si ce domaine k est réutilisé dans une séquence j.

Une approche simple de type glouton a été choisie. Cette approche considère qu'à l'étape i les résultats des étapes 1 à i-1 ne sont pas remis en cause. Cette approche procède séquentiellement sur la liste des séquences, propageant à chaque étape les domaines choisis. Pour chaque séquence, l'approche regarde le coût minimum de la fonction suivante :

$$\sum_{Occurrences} DistancesInterOccurrences + \sum_{Domaines} Taillesdesdomaines$$

Puisque le choix des domaines fait dans la première séquence est crucial pour le choix des domaines des autres séquences, une autre étape de minimisation est réalisée : chacune des séquences est choisie comme "première séquence" et le recouvrement de domaines qui possède le coût total le plus faible est conservé.

Après l'optimisation du recouvrement des séquences par les domaines, une classification finale des séquences est réalisée avec CHAVL (Figure 3.22). CHAVL [Lerman 1993] est un logiciel soutenant une méthode de classification hiérarchique de données appelée la méthode AVL (Algorithme de Vraisemblance du Lien) [127]. La classification est réalisée sur une table d'incidence de données dont les entrées sont des nombres entiers positifs (présence ou nombre d'occurrences d'un domaine dans une séquence) ou 0 (absence d'un domaine) (Figure 3.22).

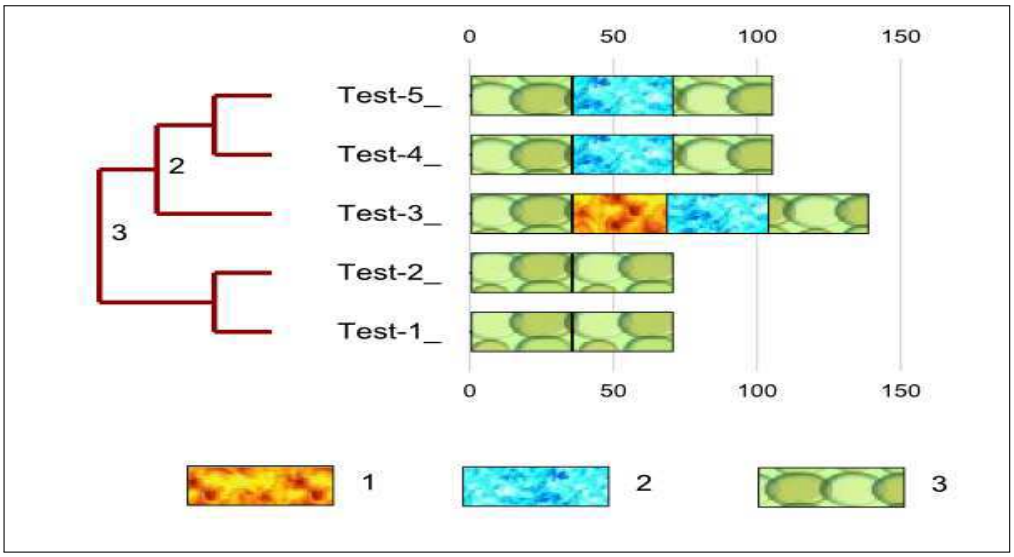


FIG. 3.26 – Exemple de visualisation et de classification de séquences synthétiques avec DomainRender. Les séquences ont été classées en fonction de la présence/absence des domaines et leur classification est représentée par l'arbre à gauche des séquences. Chaque texture correspond à un domaine distinct.

Un nouvel outil, DomainRender, a été développé en collaboration avec Mathieu Giraud, pour visualiser l'ensemble classifié de séquences avec leurs domaines (Figure 3.22 et 3.26). Les domaines sont affichés dans l'ordre gauche droite. Le programme est écrit en Python (www.python.org/) et emploie une version modifiée d'une bibliothèque de dessin créée par Baart et de Wit (www2.sfk.nl/svg/). DomainRender a deux fonctions originales : il peut importer la structure arborescente (représentation polonaise de l'arbre) de CHAVL [128] pour classer les séquences et peut afficher cet arbre à côté de la vue des domaines (Figure 3.26). Les fichiers de sorties de DomainRender sont dans le format de SVG (Scalar Vector Graphics) (www.w3.org/Graphics/SVG/), qui est un langage XML (xmlfr.org/).

Le fichier d'entrée de DomainRender est de la forme (un espace doit être présent à la fin de chaque ligne) :

```
>Nom_séquence Longueur_séquence
domain_A Position_début-Position-fin domain_B
Position_début-Position-fin ...
```

L'outil entier est utilisable via la plate-forme bioinformatique de la OUEST-genopole (www.irisa.fr/symbiose/DomainOrganizer).

3.4.2 Avantages et inconvénients de DomainOrganizer

La principale limite de la méthode est la nécessité d'un alignement multiple, effectué par ClustalW ou DIALIGN, où le gap formé par l'alignement multiple n'est pas toujours optimal. Quelques séquences, qui ne possèdent pas de domaines insérés mais ont une similitude avec une partie des domaines insérés, sont ainsi dédoublées en un certain nombre de fragments de domaines. Cette similitude est le résultat d'un nombre restreint de positions mal appariées dans l'alignement. Dans ce cas, DomainDetector tend à réduire les domaines en fragments.

Le réglage des paramètres de DomainOrganizer doit également être réalisé avec soin. Bien que seulement deux paramètres soient utilisés, il est clair qu'ils influencent le nombre et la taille de domaines détectés. Si la taille des domaines minimale est trop petite, le nombre de domaines peut être simplement trop grand pour donner une abstraction intéressante des séquences. Au contraire, si la taille des domaines est trop grande, le nombre de domaines peut être trop limité pour formuler une interprétation biologique appropriée. Une voie intéressante de recherche serait d'ajuster la valeur de ce paramètre au comportement global de la distribution des domaines. L'autre paramètre important est le seuil d'erreur. Par défaut, ce seuil d'erreur est fixé à 25 % de la taille minimale d'un domaine, mais quelques familles d'ADN répétées peuvent exiger un seuil différent.

Néanmoins, l'étape d'optimisation permet d'augmenter la robustesse des résultats vis-à-vis de la variation des paramètres de taille et de pourcentage d'erreurs des domaines. Nous avons fait varier ces deux paramètres sur des données artificielles et la famille AtREP21. Les visualisations des domaines pour les données artificielles ont montré la même organisation de domaines pour des tailles minimales de domaines qui variaient de 10 à 20 nucléotides et un pourcentage d'erreurs variant de 10 à 25 %. Pour la famille AtREP21, nous avons fait varier le paramètre de taille de 20 à 35 nucléotides et le paramètre d'erreurs de 15 à 35 %. A l'exception des tests avec le paramètre $MinSizeDomain = 35$, les différentes copies des AtREP21 sont réparties dans les mêmes sous-groupes créés par CHAVL. La différence de répartition des séquences avec une taille minimale de domaine à 35 est due à la disparition de domaines distincts observés dans les autres jeux de tests. La comparaison de la visualisation des autres tests montre de légères variations dans l'organisation en domaines des séquences. Ces variations observées sont majoritairement des différences dans les bordures de domaines ou des fusions de domaines.

L'unique algorithme comparable est MOSAIC [6], qui emploie également un alignement de ClustalW et le calcul de l'index *pi* pour la segmentation des séquences. Notre méthode de segmentation diffère par au moins trois points principaux :

- notre méthode utilise deux paramètres, tandis que MOSAIC en utilise quatre (fréquence, seuil de distance, $MinSize$ des domaines et fréquence relative).

- MOSAIC ne produit pas une vraie segmentation des séquences mais plutôt une identification des parties homogènes de l'alignement. En revanche, notre approche inclut une étape d'identification et une étape d'optimisation qui mène à une description de chaque séquence par les domaines.
- MOSAIC ne recherche pas des domaines répétés puisqu'il n'essaie aucune caractérisation des segments.

Un autre avantage de DomainOrganizer est sa grande modularité. La plupart des outils utilisés peuvent être échangés si nécessaire. La caractérisation des domaines peut être fournie par exemple par des algorithmes combinatoires de découverte de motifs tels que PRATT [90], SMILE ou RISOTTO [136]. Cependant, la découverte explicite de motifs demeure d'avantage adaptée à la découverte des motifs dans les séquences protéiques. Les HMMs se sont avérés efficaces sur la question de la segmentation de l'ADN [161], mais ont été employés plutôt dans le but de modéliser des séquences entières, ce qui limite dans la pratique la complexité des modèles. Ainsi, HMMER a été appliqué en combinaison avec RepeatMasker pour l'identification globale des ETs [92], ce qui a mené à des problèmes combinatoires et n'a pas permis de retrouvé l'architecture fine en domaines de chaque ET. L'analyse phylogénétique peut remplacer notre méthode de classification. Cependant, les méthodes emploient souvent tous les nucléotides plutôt que des domaines, et l'index dans notre méthode offre une discrimination plus fine des séquences. Par exemple, dans le cas d'une table de données T, croisant un ensemble d'individus et un ensemble d'attributs binaires, considérons la matrice S associant à chaque paire x, y d'individus le nombre de $s(x, y)$ des attributs communs à x et à y. Si la phylogénie parfaite peut être dérivée de T, alors la matrice S est ultramétrique et peut être regardée par l'intermédiaire d'un arbre ultramétrique de classification. Les méthodes de parcimonie travaillent sur des états de transformation d'une séquence ordonnée de caractères dans le but de réduire au minimum le nombre de transformations. Avec ce processus, l'index de similarité pris en considération est $s(x, y)$. Cet index est considéré dans CHAVL [128] comme un index brut qui produit un index statistiquement normalisé par rapport à sa distribution empirique sur des paires d'individus. De cette façon, les effets non significatifs sont neutralisés. En outre, la version finale de l'index utilisé pour l'agrégation se rapporte à une échelle probabiliste, c'est-à-dire à la probabilité d'observer la valeur de l'index précédent. De façon générale, cette méthode prouve qu'une famille de séquences d'ADN répétés peut être décrite et classifiée à un niveau macroscopique par des domaines. CHAVL, qui emploie la matrice de distribution des domaines pour classifier les séquences, semble donc plus appropriée pour cette méthode d'analyse.

Résumé

Une étude préliminaire des hélitrons non-autonomes avait montré que ces éléments n'avaient pas une séquence interne aléatoire mais une séquence interne organisée en domaines nucléaires. Nous avons créé un logiciel nommé DomainOrganizer qui permet la visualisation et la classification des séquences en fonction de leur composition en domaines. Ce logiciel est composé de plusieurs algorithmes originaux dont DomainDetector, DomainOptimizer et DomainRender. DomainDetector segmente un alignement multiple en domaines. DomainOptimizer minimise le nombre de domaines nécessaires pour couvrir l'ensemble des séquences. Enfin, DomainRender lit l'arbre de classification et la position des domaines, puis crée le fichier SVG de visualisation des séquences organisées en

domaines. Néanmoins, la segmentation dépend des valeurs des paramètres passés en arguments et ses valeurs sont variables d'une famille à l'autre. De plus, le découpage des séquences en segments dépend initialement de l'alignement multiple qui n'est pas toujours performant pour des séquences hautement divergentes. Une amélioration majeure de DomainOrganizer serait de s'affranchir de l'alignement multiple et d'utiliser une méthode de découpage optimale et non dépendante de la taille des domaines.

Contribution scientifique

Article : Tempel S., Lerman I.C., Giraud M., Valin A.S., Couée I., El Amrani and A. Nicolas J. 2006. Domain Organization within repeated DNA sequences : Application to the study of a new family of transposable elements in Arabidopsis thaliana. Bioinformatics, 22 :1948-1954

Conférence : Tempel S., Lerman I.C., Giraud M., Valin A.S., Couée I., El Amrani and A. Nicolas J. 2004. Genome-wide analysis of domain organization in Arabidopsis helitronic structures. XII ème Colloque Eléments Transposables. Tours.

Conférence : Veber P., Tempel S., Andonov R., Lavenier D. and Nicolas J. Roadef 2007. Détection de domaines dans des séquences génomiques : un problème de couverture optimale. 2007. Grenoble.

3.5 Mise en évidence de l'organisation en domaines de la famille d'hélitron AtREP21

3.5.1 Test de DomainOrganizer sur une famille de MITE

L'outil DomainOrganizer [196] a d'abord été testé sur des séquences artificielles répétées et sur des éléments transposables non-autonomes tels que le MITE Emigrant (Figure 3.27).

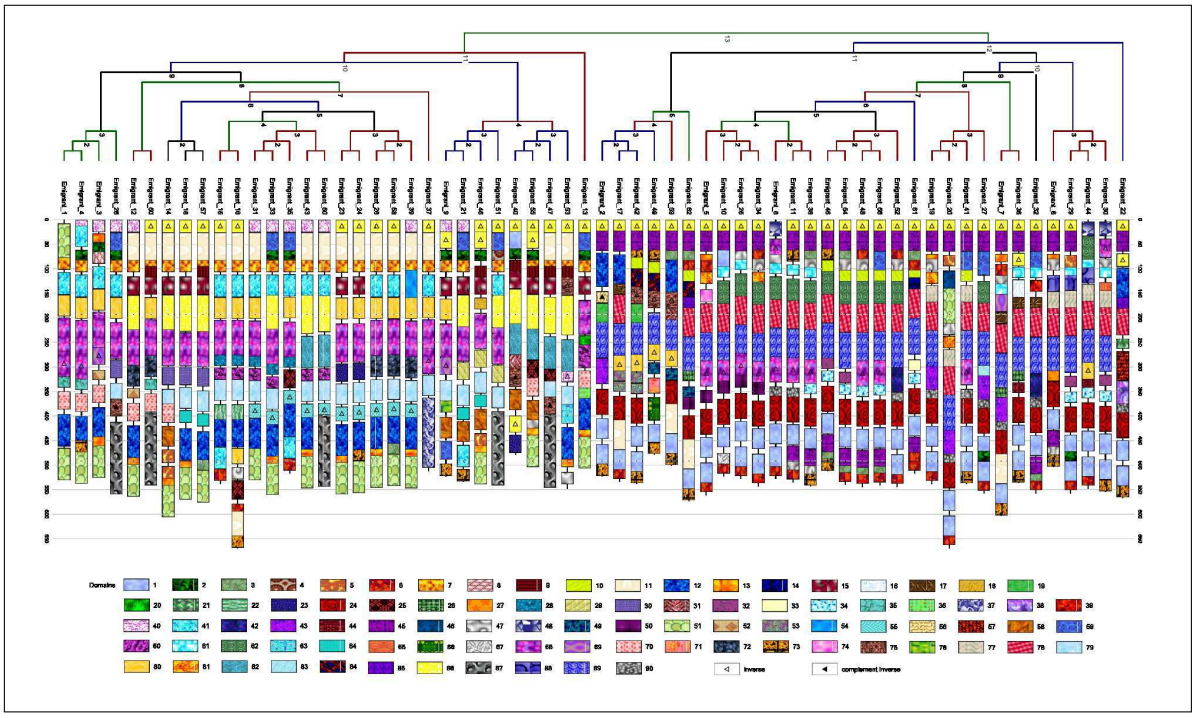


FIG. 3.27 – Test de DomainOrganizer sur l'organisation en domaines de la famille Emigrant [69]. L'image obtenue par DomainRender a été tournée horizontalement. Nous observons que les extrémités 3' des deux sous-groupes d'Emigrants, qui possèdent la même séquence (25 pb), sont représentées comme des domaines distincts. Cette différence est due à l'alignement multiple. Les deux sous-groupes n'ont pas été alignés sur ces extrémités et DomainDetector les a considéré comme deux domaines différents.

Les séquences de cette famille ont été récupérées avec STAN [154]. Les séquences des TIRs et les TSD ont été récupérés dans l'article de Feschotte et Mouchès [69]. Comme les hélitrons non-autonomes, les MITEs ont une séquence interne très variable et sont recopiés grâce à la reconnaissance de leurs extrémités. Nous avons donc utilisé le même modèle que le modèle utilisé pour la recherche des hélitrons : une extrémité 5' (TIR + TSD), une séquence quelconque de taille bornée et une extrémité 3' (TIR + TSD). Les TIRs et TSD ont été recherchés avec une erreur de substitution et séparés par une séquence quelconque de 0 à 1000 pb.

La Figure 3.27 montre la composition en domaines de la famille Emigrant chez *Arabidopsis thaliana*. A l'exception de l'Emigrant numéro 13, les copies Emigrants sont divisées en deux sous-groupes. Les deux sous-groupes sont très distincts en composition de domaines, mais les copies présentes dans chaque sous-groupe sont similaires en composition

de domaines. A partir de ce résultat, nous avons pu suggérer deux hypothèses distinctes : soit les copies Emigrants ont été créées à partir du même transposon autonome, mais à partir de deux événements différents de transposition ; soit elles ont été créées après la copie d'un premier Emigrant, cette copie a subi de nombreuses mutations pour donner une nouvelle composition en domaines internes. Nous pensons que la deuxième hypothèse est peu vraisemblable, car elle sous-entend que la copie de l'Emigrant a été entièrement mutée et qu'elle a conservé sa taille d'origine (Figure 3.27). Les différents travaux sur les MITEs tendent, au contraire, à prouver que les éléments transposables non-autonomes mutent par délétions de leur séquence interne [39, 89].

Si la première hypothèse est exacte, les séquences MITEs, définies comme appartenant à la famille Emigrant, devraient être considérées comme deux familles distinctes. La première famille est composée des Emigrants 2, 5-8, 10-11, 17, 19, 20, 22, 25, 27, 29-30, 32, 34, 36, 38, 41-42, 44, 46, 48-49, 51, 59, 61-62, 64, 66. La deuxième famille est composée des Emigrants 1, 3-4, 9, 12-16, 18, 21, 23-24, 26, 28, 31, 33, 35, 37, 39-40, 43, 45, 47, 50-51, 53, 55, 57, 60, 68.

3.5.2 Détection des domaines de la famille AtREP21

Nous avons ensuite testé notre méthode sur une famille d'hélitrons non-autonomes pour essayer de comprendre le mécanisme de variabilité de ces éléments et essayer de retracer l'histoire évolutive de cette famille.

Une étude préliminaire de la famille non-autonome AtREP3 a permis de découvrir la présence de deux nouvelles familles d'hélitrons insérées dans les séquences internes de deux copies d'AtREP3 : AtREP20 et AtREP21 (Figure 3.5). Nous avons choisi d'étudier la famille AtREP21 pour comprendre la création et l'organisation de la séquence interne des hélitrons non-autonomes. Nous avons réutilisé le modèle syntaxique de détection des hélitrons avec la famille AtREP21 et obtenu ce modèle :

```
TCCCTTTATTATTAAAAGGGAAGTACAAATTGAAAT:9
-x(100,2500)-
CCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTAG:9
```

Cette famille a 48 séquences AtREP21 dans le génome d'*Arabidopsis thaliana* et ces séquences mesurent de 315 à 1012 nucléotides. Nous avons commencé par détecter et visualiser l'ensemble des domaines présents dans les séquences de la famille AtREP21. Cette visualisation a permis d'identifier des groupes d'AtREP21 et de comprendre certains mécanismes biologiques de variabilité de ces séquences répétées. Puis, nous avons identifié les domaines internes pour retracer l'histoire évolutive de cette famille.

3.5.3 Organisation en domaines de la famille AtREP21

Les paramètres d'alignement de ClustalW étaient de 25 pour l'ouverture du gap et de 0.01 pour l'extension du gap. La taille choisie pour MinSizeDomain était de 26 pb avec une valeur de $MaxErrors$ égale à 25 %. L'alignement multiple est répertorié dans l'annexe 5.2 Le seuil d'acceptation de la E-value pour fusionner les domaines a été de 10^{-5}. DomainDetector (Figure 3.28) a détecté 37 frontières. Il y avait 121 domaines présents avant l'étape d'optimisation des domaines et 60 domaines après cette optimisation. Les

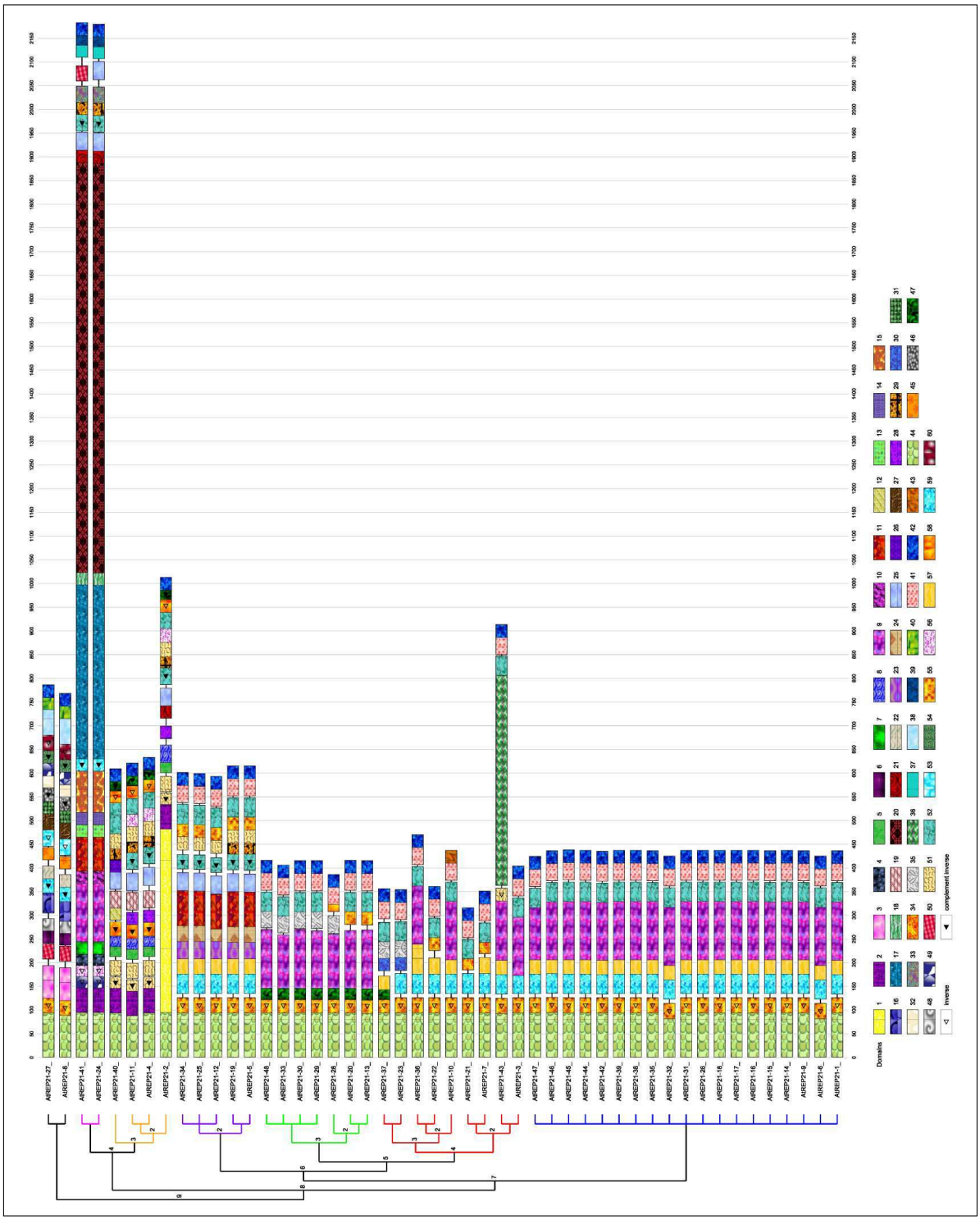


FIG. 3.28 – Organisation en domaines de la famille d'hélitron AtREP21 [196]. Chaque domaine est identifié par une texture unique attribuée par DomainRender.

séquences de ces domaines sont répertoriées dans l'annexe 5.3. Les extrémités 5' et 3', caractéristiques de la famille AtREP21, sont respectivement décrites par les domaines 44 et 42.

Comme les autres AtREPs [98], la famille AtREP21 montre une grande variation dans ses séquences internes (Figure 3.22). Ces variations de séquences résultent principalement d'insertions/délétions de domaines. La visualisation créée par DomainRender (Figure 3.22) montre que les AtREP21 sont divisés principalement en sept groupes de séquences.

Le premier groupe (couleur bleue Figure 3.28) est composé des AtREP21 1, 6, 9, 14-18, 26, 31-32, 35, 38-39, 42, 44-47. Ce groupe est principalement composé d'une combinaison des domaines 9, 42, 44, 55-i (inversé), 52, 56-57 et 59. Ce groupe est l'ensemble des séquences les plus conservées d'AtREP21. Les séquences ont la même combinaison de domaines et une taille similaire comprise entre 400 et 450 pb (Figure 3.28). Nous avons comparé ensuite les autres groupes à ce premier groupe en termes de domaines.

Le deuxième groupe (couleur rouge) contient les AtREP21 3, 7, 10, 21-23, 36-37 et 43. Il est principalement caractérisé par la délétion du domaine 9 et/ou la délétion/substitution ponctuelle d'un domaine présent dans le groupe 1. L'AtREP21 43 contient une large insertion du domaine 36.

Le troisième ensemble (couleur verte) est composé des AtREP21 13, 20, 28-30, 33 et 48. Il est caractérisé par la substitution (par rapport au premier groupe) des domaines 57 et 59 par le domaine 47 et par l'insertion du domaine 35 ou du domaine 55.

Le quatrième groupe (couleur violette, Figure 3.28) est composé des AtREP21 5, 12, 19, 25 et 34. Il est identifié grâce à la large substitution du domaine 9 par les domaines 11, 23-25, 51, 52-ci (complémentaire-inversé) et 55. A l'exception des AtREP21 5 et 19 qui contiennent l'insertion supplémentaire du domaine 29, ce groupe est aussi très similaire, avec la même combinaison de domaines pour ses copies d'AtREP21.

Le cinquième groupe (couleur rouge) est défini par les AtREP21 2, 4, 11 et 40, et partage avec les précédents groupes seulement les domaines 42, 44 (les domaines spécifiques des AtREP21) et 52. A l'exception de l'AtREP21 2 qui comprend le domaine supplémentaire noté 1, le groupe est défini par les domaines 5-6, 8, 19, 25, 28-29, 42, 44, 47, 51, 51-52-ci, 56, 58-ci et 58-i.

Le sixième groupe est composé des AtREP21 24 et 41. Ce petit groupe est composé des domaines 4, 7-11, 13-15, 17-18, 20-21, 25, 29, 33, 37, 39, 42, 44, 49-i, 52-ci, 56-i et 59-ci.

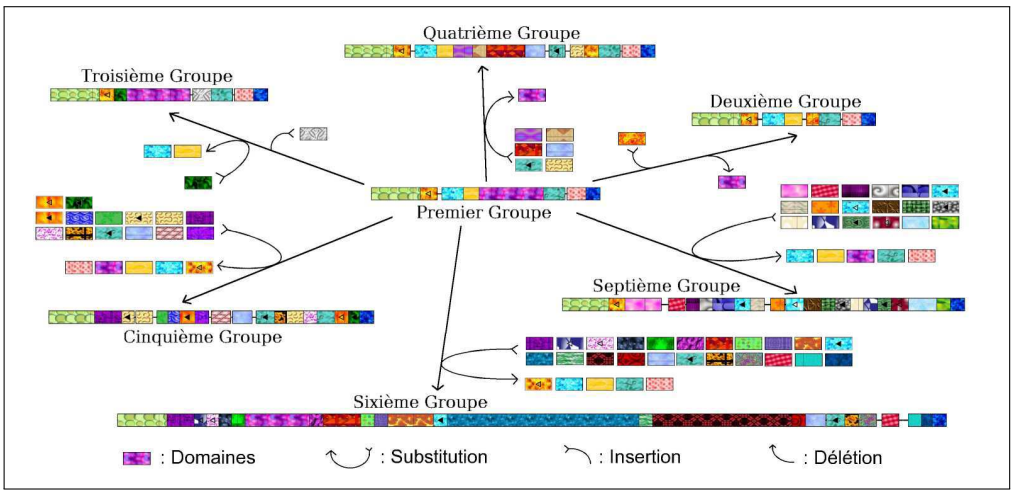


FIG. 3.29 – Comparaison entre le premier sous-groupe d'AtREP21 et les autres sous-groupes. A chaque comparaison est indiquée les domaines insérés, délétés ou substitués.

Le dernier groupe est composé des AtREP21 8 et 27. Ce groupe contient en plus des domaines 42 et 44, les domaines 3, 6, 16, 22, 27, 31-32, 38, 40, 45, 46-ci, 48-50, 53-i, 54-ci, 55-i, 59-ci et 60-i.

Notre méthode DomainOrganizer [196] a révélé la complexité et la variation des domaines internes entre les différentes copies de la famille AtREP21 (Figure 3.29), que des méthodes standard telles que BLAST ou PILER ne peuvent observer.

3.5.4 Domaines internes et structures secondaires

Nous avons voulu savoir si les domaines détectés dans la famille AtREP21 étaient des "vrais" domaines ou bien étaient des artefacts créés par notre logiciel. S'il s'agit de domaines "vrais", il est probable qu'ils aient un rôle biologique pour l'élément AtREP21. Ces hélitrons sont non-autonomes et la recherche d'ORF dans cette famille a déjà été réalisée sans succès dans la section 3.3.3. Nous avons alors recherché un rôle structurel de ces domaines pour la famille AtREP21. Nous avons utilisé le logiciel MFold 3.2 (www.bioinfo.rpi.edu/applications/mfold/) [213] pour connaître la structure secondaire de tous les AtREP21 (Figure 3.30), puis de tous les domaines (Figure 3.30).

Nous avons observé que la plupart des domaines, créés par DomainOrganizer, ne possèdent pas de structure secondaire bien définie par MFold, à l'intérieur d'une copie d'AtREP21. Par exemple, la structure secondaire du domaine 9 varie d'une copie d'AtREP21 à une autre. Il est donc très difficile d'attribuer un rôle structurel particulier à un domaine donné. Néanmoins, la comparaison de deux hélitrons, ayant une organisation en domaines similaires, a montré que la délétion, l'insertion ou la substitution d'un (ou plusieurs) domaine(s) modifie la structure secondaire des hélitrons. Par exemple, l'insertion du domaine 35 dans l'AtREP21 29 modifie significativement la structure secondaire de la séquence interne par rapport à la structure secondaire de l'AtREP21 45 (Figure 3.30). Nous pouvons donc supposer que la variation de la séquence interne des hélitrons est due à la modification de la structure secondaire des hélitrons.

Nous avons aussi observé la structure secondaire de chaque domaine isolé des AtREP-21. Chaque occurrence d'un domaine donné est alignée avec ClustalW [200], et cet alignement est utilisé pour créer le profil HMM de chaque domaine. Le profil est ensuite converti en une séquence consensus qui est utilisée avec MFold 3.2.

La plupart des domaines n'ont pas de structure secondaire bien définie telle que des hairpins ou des palindromes. Seules les insertions de grands domaines et quelques domaines, tels que les domaines 16 et 24, présentent des structures secondaires stables. Les structures stables des grandes insertions semblent montrer que ces domaines sont des éléments transposables, même si leur identification n'a pas toujours été possible. Le manque de structure a plusieurs causes possibles : le domaine est un artefact, les bordures du domaine sont fausses et/ou le domaine ne forme une structure qu'avec un autre domaine. Il est possible que les paramètres choisis pour la segmentation des domaines créent des subdivisions artificielles de domaines. Nous observons alors des parties de domaines qui n'ont aucune structure secondaire cohérente. Lors d'une grande insertion ou délétion de séquence, un alignement multiple aura des difficultés à délimiter cette séquence. Ainsi, DomainOrganizer, qui lit l'alignement, choisit la mauvaise limite du domaine. Enfin, un domaine détecté peut être la séquence palindromique d'un autre domaine et ils forment ensemble une structure en tige-boucle (Figure 3.30). Ainsi, la recherche de structures secondaires de l'un de ces deux domaines palindromiques, sans l'autre domaine, ne montre

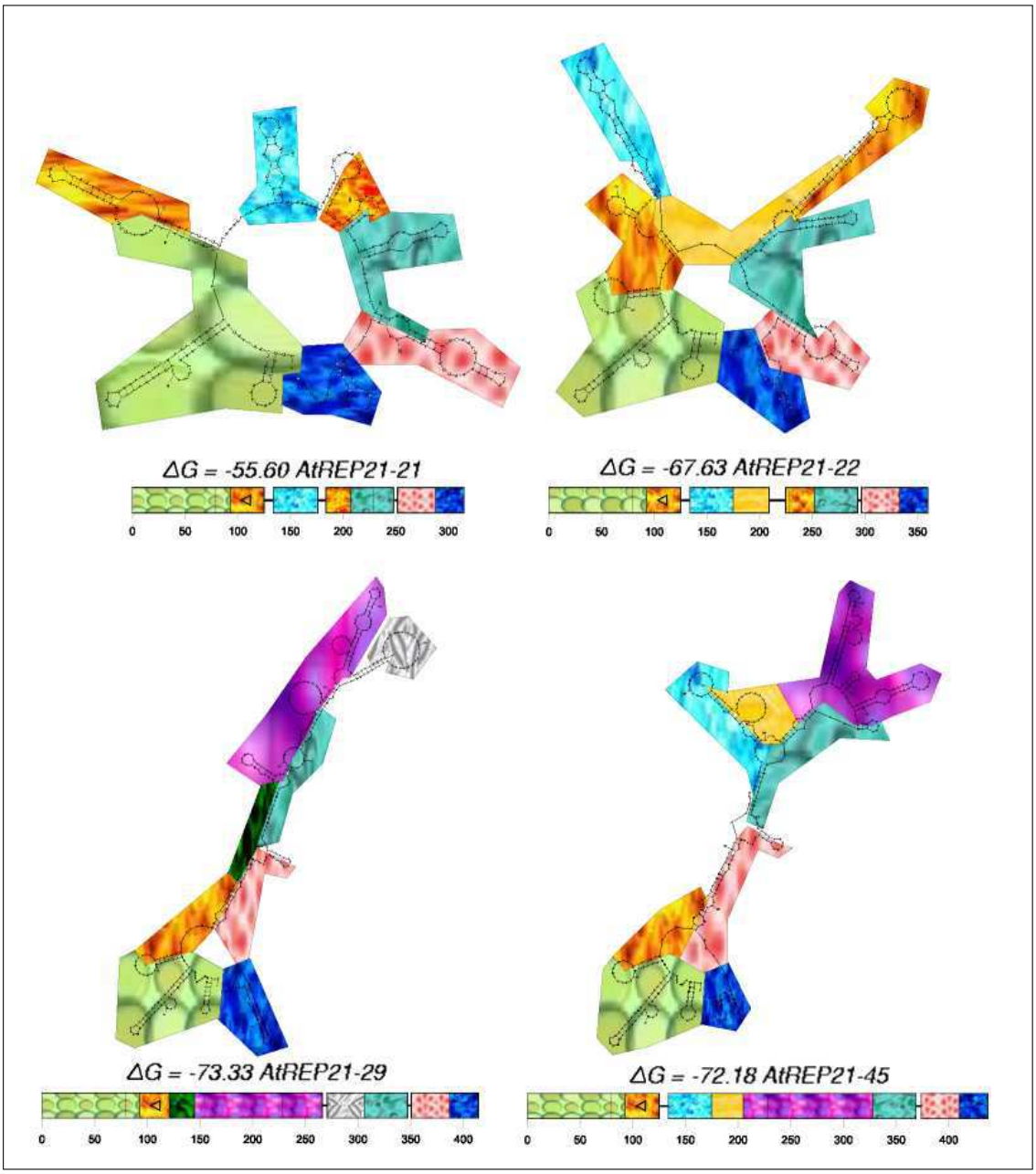


FIG. 3.30 – Exemple de structures secondaires associées à la segmentation en domaines. Dans cet exemple nous comparons l'AtREP21 21 avec l'AtREP21 22 et l'AtREP21 29 avec l'AtREP21 45. Nous avons mis la texture des domaines trouvé avec DomainOrganizer sur les structures secondaires. La valeur donnée par ΔG représente la stabilité de la séquence à conserver cette structure.

pas de structure.

Le domaine 42 (extrémité 3' des AtREP21) possède normalement l'hairpin caractéristique des hélitrons (Figure 3.30 et 3.31). Curieusement, le domaine 44 (extrémité 5'), qui d'après la littérature ne présente aucune caractéristique connue [98], montre avec MFold une structure secondaire stable. Malheureusement, aucune fonction n'a été trouvée et cette structure secondaire n'a pas été retrouvée dans les toutes autres familles d'hélitrons.

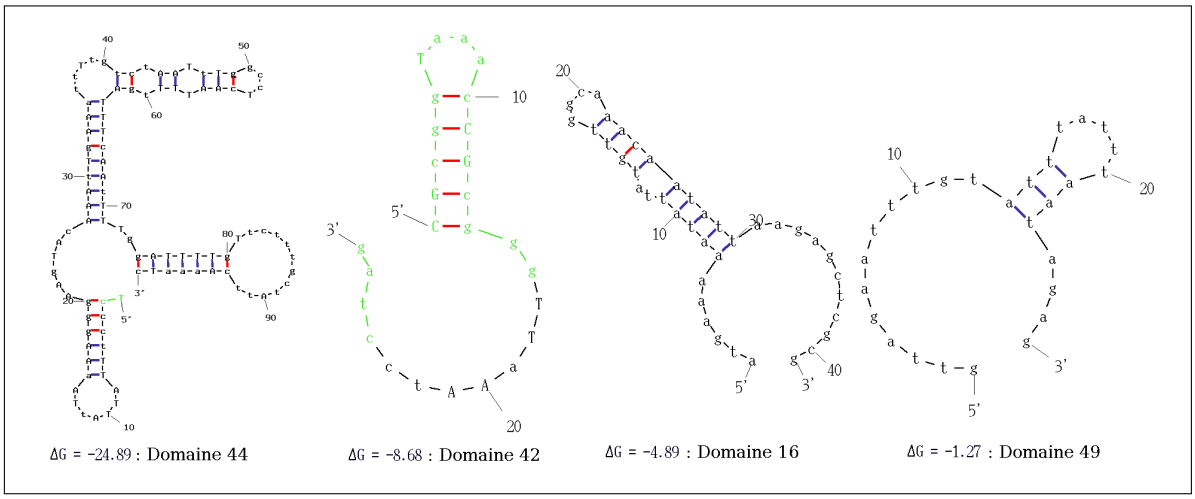


FIG. 3.31 – Structure secondaire associée à quatres domaines. Les domaines présentés sont les domaines 42 et 44, caractéristiques des extrémités AtREP21, et les domaines 16 et 49. Les nucléotides colorés en vert représente les nucléotides caractéristiques des hélitrons [98]. Le domaine 16 est un exemple de domaine ayant une structure secondaire bien définie (une tige-boucle). Le domaine 49 ne présente aucune structure secondaire. Les traits rouges représentent les liaisons CG et les traits bleus représentent les liaisons AT.

Il serait intéressant de vérifier l'importance de cette structure, par exemple, en mesurant *in vivo* le taux de transposition de cette famille en fonction de la dégradation de cette structure secondaire.

3.5.5 Identification des domaines internes d'AtREP21

A l'exception de quelques domaines tels que le domaine 1 qui semble être un domaine inséré (Figure 3.28) et qui présente des extrémités hélitroniques (domaine identifié comme étant un hélitron3), la visualisation des domaines n'explique pas la nature et/ou les mécanismes biologiques de variation des domaines internes.

Pour connaître la nature des domaines de la famille AtREP21 et comprendre l'évolution de cette famille, nous avons compté les occurrences de chaque domaine et leurs proportions à l'intérieur/extérieur de la famille AtREP21. Ensuite, nous avons observé leurs répartitions dans le génome de chaque domaine ou groupe de domaines. Enfin, nous avons cherché dans leurs séquences des motifs caractéristiques permettant leur identification.

3.5.5.1 Distribution des domaines internes dans *Arabidopsis thaliana*

Le nombre d'occurrences de chaque domaine est un bon indicateur de la nature biologique du domaine. Par exemple, si un domaine est présent en plusieurs centaines de copies et qu'AtREP21 est présent en 48 copies, cela montre que le domaine a son propre mécanisme d'invasion du génome (éléments transposables ou mini/microsatellites).

Pour chaque domaine interne, à partir de toutes ses positions dans les copies d'AtREP21, les séquences sont récupérées et un profil HMM est créé par HMMER [59]. Deux approches différentes sont alors utilisées pour rechercher ce domaine. Dans la première,

le profil est directement recherché dans le génome par l'outil HMMsearch [59]. Nous ne conservons que les matches ayant une E-Value inférieure à 10^{-4}. Dans la deuxième approche, le profil est transformé en séquence consensus qui est recherchée par STAN [154]. Dans ce cas, si le domaine mesure moins de 40 nucléotides, il est directement utilisé dans STAN avec 25 % d'erreurs de substitutions, sinon les 20 premiers et les 20 derniers nucléotides sont conservés avec 25 % d'erreurs et un gap flexible est inséré entre ces deux extrémités. La taille de ce gap est calculée grâce à la taille mesurée des différentes copies de ce domaine.

Nous avons choisi de conserver ces deux méthodes car une étude préliminaire, sur plusieurs domaines internes, a montré que ces deux méthodes sont complémentaires. Elles permettent d'éliminer les faux positifs (occurrences présentes mais ne correspondant pas à un domaine) obtenus par chacune des méthodes. STAN détecte bien les grands domaines. Les plus petits domaines ne sont pas assez restrictifs à l'échelle du génome et STAN détecte alors des fausses occurrences pour ces domaines. HMMER détecte mal dans le génome d'*Arabidopsis thaliana* les domaines ayant une faible complexité de séquence (par exemple des domaines uniquement composés par les nucléotides A et T).

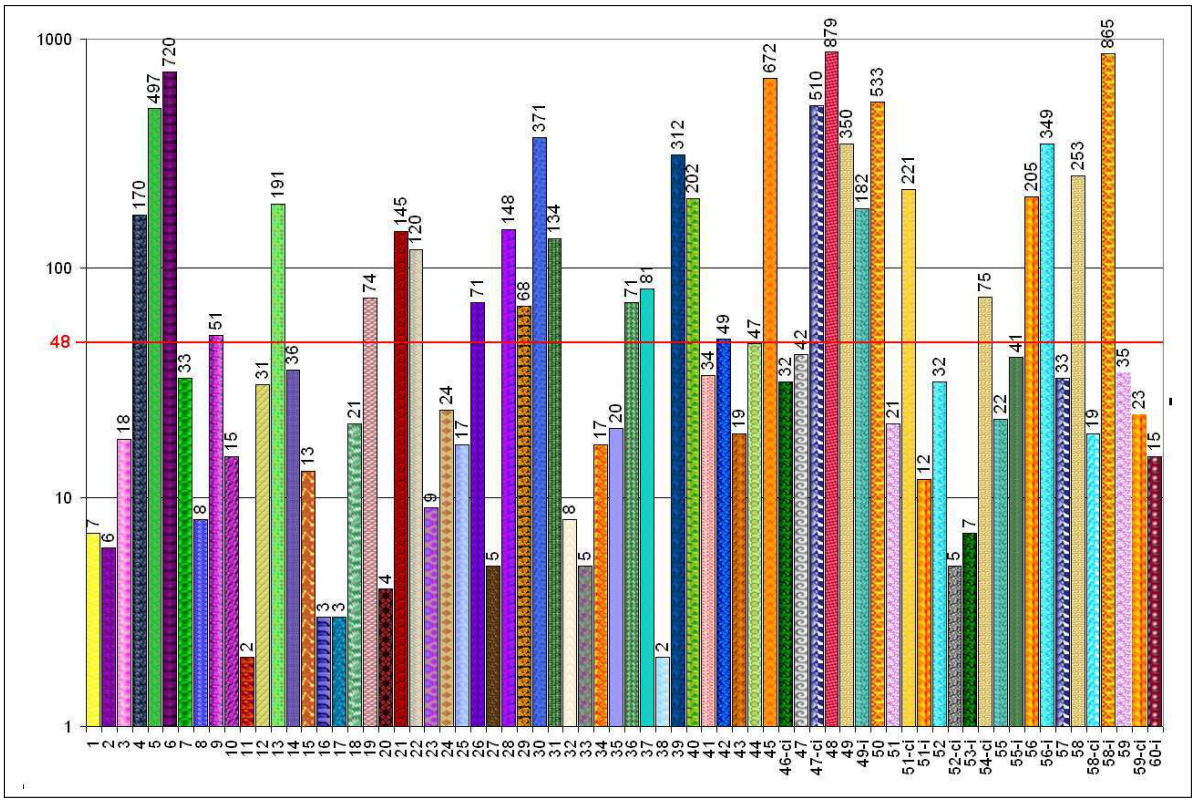


FIG. 3.32 – Nombre d'occurrences de chaque domaine interne d'AtREP21 présent dans le génome d'*Arabidopsis thaliana*. Les textures obtenues avec DomainRender pour la famille AtREP21 sont reportées pour chaque domaine. Le seuil de 48 copies est indiqué en rouge, il délimite les domaines qui sont plus nombreux que les copies d'AtREP21 dans le génome.

Nous avons alors comparé les positions respectives de chaque occurrence. Nous n'avons conservé que les occurrences ayant les mêmes positions (ou chevauchantes) pour les deux

méthodes (Figure 3.32). Curieusement, certains domaines ont des valeurs d'occurrences inférieures aux valeurs observées par DomainRender. Par exemple, le domaine 10 a sept occurrences dans la Figure 3.28 et alors que nous n'avons retrouvé que trois occurrences par l'association de ces deux méthodes. Cette différence de valeur est due à l'élimination de faux négatifs (vraies occurrences qui ne sont pas détectées). Néanmoins, toutes les occurrences trouvées sont de vrais occurrences de domaines.

Curieusement, pour certains domaines, cette méthode a aussi compté plus d'occurrences dans les 48 copies d'AtREP21 que d'occurrences que l'on peut compter en observant la Figure 3.28. Par exemple, le domaine 19 n'est présent que trois fois sur la Figure 3.28, mais est compté sept fois par notre méthode. Cette différence s'explique, car la Figure 3.28 ne représente que le résultat de l'optimisation des domaines sur les copies d'AtREP21 et non toutes les occurrences de chaque domaine : ainsi les quatre occurrences manquantes du domaine 19 sont des occurrences qui n'ont pas été sélectionnées par l'optimisation de DomainOptimizer.

Les fréquences d'occurrences des domaines sont très variables : le domaine 11 a seulement deux occurrences alors que le domaine 45 a 672 occurrences. Le nombre d'AtREP21 présents dans le génome d'*Arabidopsis thaliana* (48 occurrences) est très inférieur au nombre d'occurrences de beaucoup de domaines (Figure 3.32). Ces hautes valeurs laissent supposer que la plupart des domaines présents dans AtREP21 ne sont pas spécifiques de cette famille, mais ont été capturés [15] ou insérés puis copiés avec les séquences d'AtREP21.

3.5.5.2 Répartition des domaines internes dans les AtREP21, les hélitrons et le reste du génome

Cette hypothèse a pu être vérifiée en comptabilisant le nombre d'occurrences présentes à l'intérieur des AtREP21 ou dans les autres hélitrons par rapport au nombre d'occurrences présentes ailleurs dans le génome. Chaque position de domaine est comparée avec les positions des AtREP21, des chimères d'AtREP21 et des autres hélitrons. Si l'occurrence est comprise entre les positions de début et de fin des AtREP21, d'une chimère ou d'un hélitron, cette occurrence est considérée comme appartenant à cet hélitron. Sinon nous avons considéré que cette occurrence était isolée des hélitrons.

La plupart des domaines ont une grande partie de leurs occurrences à l'extérieur des AtREP21 et des hélitrons : 42 domaines (60 au total) ont plus de la moitié de leurs occurrences à l'extérieur de tous les hélitrons connus et 23 domaines, parmi ces 42 domaines, ont plus de 90 % de ces occurrences en dehors des hélitrons (Figure 3.33). A l'exception du domaine 56 qui a 43 % de ses occurrences liées aux hélitrons, tous les domaines qui ont plus de 100 occurrences dans le génome entier d'*Arabidopsis thaliana* sont des domaines très faiblement associés aux hélitrons (moins de 20 % d'occurrences dans les AtREP21 et les hélitrons).

Ces résultats montrent une forte corrélation négative entre la fréquence d'apparition des domaines et leur spécificité par rapport à AtREP21 : plus les domaines sont liés aux AtREP21, moins leur occurrence est grande. Par exemple, le domaine 48 a 879 occurrences, dont moins de 8 % présentes dans les hélitrons (AtREP21 inclus), et le domaine 11 a 2 occurrences dont 100 % sont contenues dans les AtREP21 (Figure 3.32 et 3.33).

A l'exception des domaines 24, 43, 56 et 60-i, tous les domaines sont peu présents dans les hélitrons autres qu'AtREP21 et ses chimères (pourcentage inférieur à 25 %) (Figure

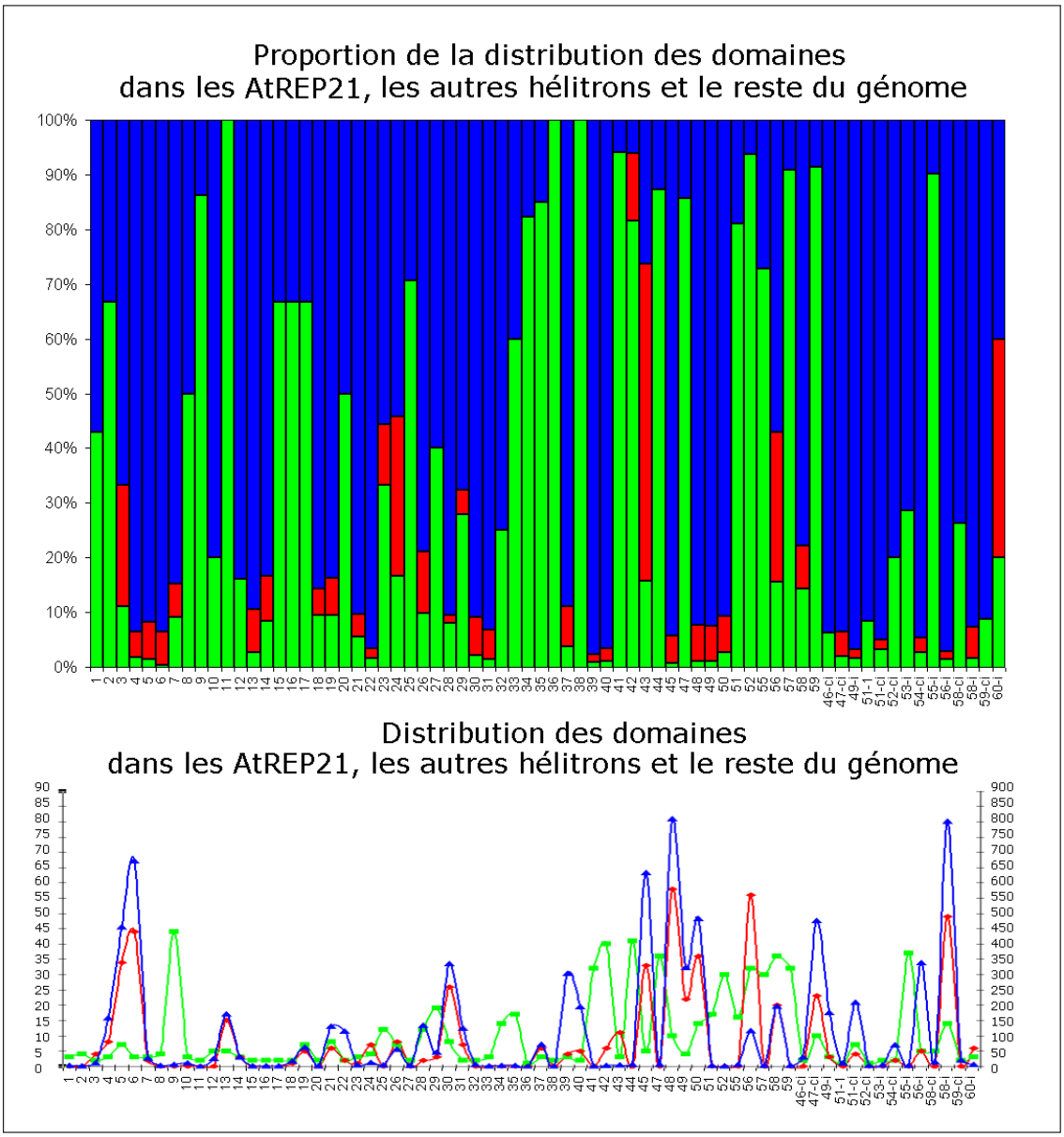


FIG. 3.33 – Pourcentage d'occurrences présentes dans les AtREP21, les hélitrons et le reste du génome. Le graphe du haut empile trois pourcentages d'occurrences des domaines internes d'AtREP21. Les barres vertes représentent le pourcentage d'occurrences de chaque domaine présent dans les AtREP21 ou dans des chimères d'AtREP21. Les barres rouges montrent le pourcentage d'occurrences présent dans les autres familles d'hélitrons. Les dernières barres (en bleu) représentent le pourcentage d'occurrences en dehors des hélitrons chez *Arabidopsis thaliana*. Les courbes du graphe du bas montrent les occurrences de chaque catégorie de domaines avec les mêmes couleurs : vert pour les AtREP21, rouge pour les hélitrons et bleu pour le reste du génome.

3.33). Le domaine 43 est uniquement présent dans l'AtREP21 10, mais il est largement présent dans les autres familles d'hélitrons (Figure 3.33). L'alignement de sa séquence avec la séquence du domaine 42 (domaine 3' caractéristique des AtREP21) a montré qu'il s'agit de la même séquence ayant subi certaines mutations. Comme les extrémités 3' des

différentes familles d'hélitrons sont similaires, il est normal de retrouver ce domaine 43 muté dans les autres hélitrons. Le domaine 56 est riche en A+T, et l'étude de ses positions dans le génome montre qu'il présente les caractéristiques d'un minisatellite. Inukai *et al.* ont démontré l'insertion de minisatellite dans les hélitrons du riz [86]. Les minisatellites utilisent les éléments transposables non-autonomes comme vecteurs pour leur réplication [85] ce qui explique sa relative proportion dans les autres hélitrons (Figure 3.33).

Ce résultat semble montrer que les séquences internes des copies d'AtREP21 se sont très peu (ou pas du tout) échangées avec les séquences internes des autres hélitrons. Nous pouvons alors supposer qu'un hélitron autonome (ou non-autonome) a créé l'AtREP21, puisque cet élément a subi des mutations à l'intérieur de ses copies, mais sans intervention d'une autre famille d'hélitrons.

Seuls deux domaines sont uniquement présents dans les AtREP21 : les domaines 11 et 38. Cinq autres domaines sont présents à plus de 90 % dans les AtREP21 : les domaines 41, 52, 55-i, 57 et 59 (Figure 3.33). Les domaines 55-i, 57 et 59 sont des domaines présents dans trois des cinq groupes d'AtREP21 et sont toujours accompagnés du domaine 44, le domaine de l'extrémité 5' d'AtREP21. De plus, les domaines 41 et 52 sont positionnés juste avant le domaine caractéristique du domaine 3' de l'hélitron. De plus, le domaine 41 contient deux des sept nucléotides qui constituent la séquence palindromique de l'hairpin subterminale d'AtREP21. Le domaine 41 avec le domaine 44 sont aussi présents dans trois groupes (Figure 3.28). Le domaine 38 est associé à l'extrémité 3' des AtREP21 8 et 27. Contrairement aux domaines précédents, le domaine 11 est au centre de la séquence interne des AtREP21. Néanmoins, l'analyse de sa séquence primaire a montré une grande similarité avec les extrémités 5' hélitroniques. Le paragraphe 3.3.5 a montré que des hélitrons peuvent être composés de plusieurs extrémités 5' ou 3'. Nous avons supposé que le domaine 11 était une extrémité hélitronique qui s'était insérée dans un AtREP21. Cette extrémité n'ayant plus de rôle fonctionnel, elle aurait dégénéré au cours du temps.

Ces résultats montrent que la famille AtREP21 possède une séquence consensus d'une centaine de nucléotides pour chacune de ses extrémités. La conservation des séquences subterminales a aussi été observée chez les MITEs (transposons non-autonomes de classe II) d'*Arabidopsis thaliana* [133]. Cette conservation des séquences subterminales indique que les processus de création des MITEs et hélitrons non-autonomes (ETs de classe II) pourraient présenter des similarités.

3.5.5.3 Nature biologique des domaines internes

Le comptage des occurrences, l'étude de leurs répartitions dans les hélitrons et l'étude de leur séquence ont permis d'identifier la nature de quelques domaines. Néanmoins, cette approche n'a pas permis d'identifier la plupart des domaines internes d'AtREP21, en particulier les domaines qui ont un fort pourcentage d'occurrences à l'extérieur des hélitrons (Figure 3.33). Une nouvelle étude des positions de chaque domaine interne d'AtREP21 a été réalisée. Les positions d'un domaine donné ont été comparées aux positions des autres domaines dans tout le génome. Deux méthodes complémentaires ont été utilisées pour comparer les positions des domaines à l'extérieur des AtREP21. La première méthode a compté les occurrences d'un domaine proches d'une autre occurrence de ce même domaine. La deuxième méthode a compté le nombre d'occurrences d'un domaine proche d'un autre domaine. Pour les deux méthodes, nous avons décidé qu'un domaine était "proche" d'un autre domaine si la distance entre les deux domaines était inférieure à 50 pb.

La première méthode permet d'identifier les différents micro/minisatellites. Un (micro/mini)satellite est une séquence répétée en tandem [21] : si l'on trouve plusieurs occurrences proches les unes des autres, on peut supposer qu'il s'agit d'un satellite. Cette approche nous a permis de caractériser les domaines 5, 6, 22, 30, 35, 41, 45, 48 et 58-i comme des minisatellites (Figure 3.34, 3.36 et 3.37). Par exemple, le domaine 35 est un domaine minisatellite dont les occurrences sont principalement liées à la famille AtREP21 (Figure 3.33). Ce résultat est analogue aux observations faites dans le génome du riz [85], les occurrences de ce minisatellite, en dehors de l'hélitron *basho*, seraient éliminées par le génome et conservées à l'intérieur des AtREP21.

La deuxième méthode permet de connaître "l'indépendance" des domaines internes. Si un domaine possède son propre mode de réplication, beaucoup d'occurrences isolées seront retrouvées. Si un domaine n'a pas de système de réplication autonome et fait partie d'une entité biologique plus grande composée de plusieurs domaines internes, ce domaine sera toujours associé avec d'autres domaines. On a comparé la position d'une occurrence d'un domaine interne avec toutes les occurrences de tous les autres domaines.

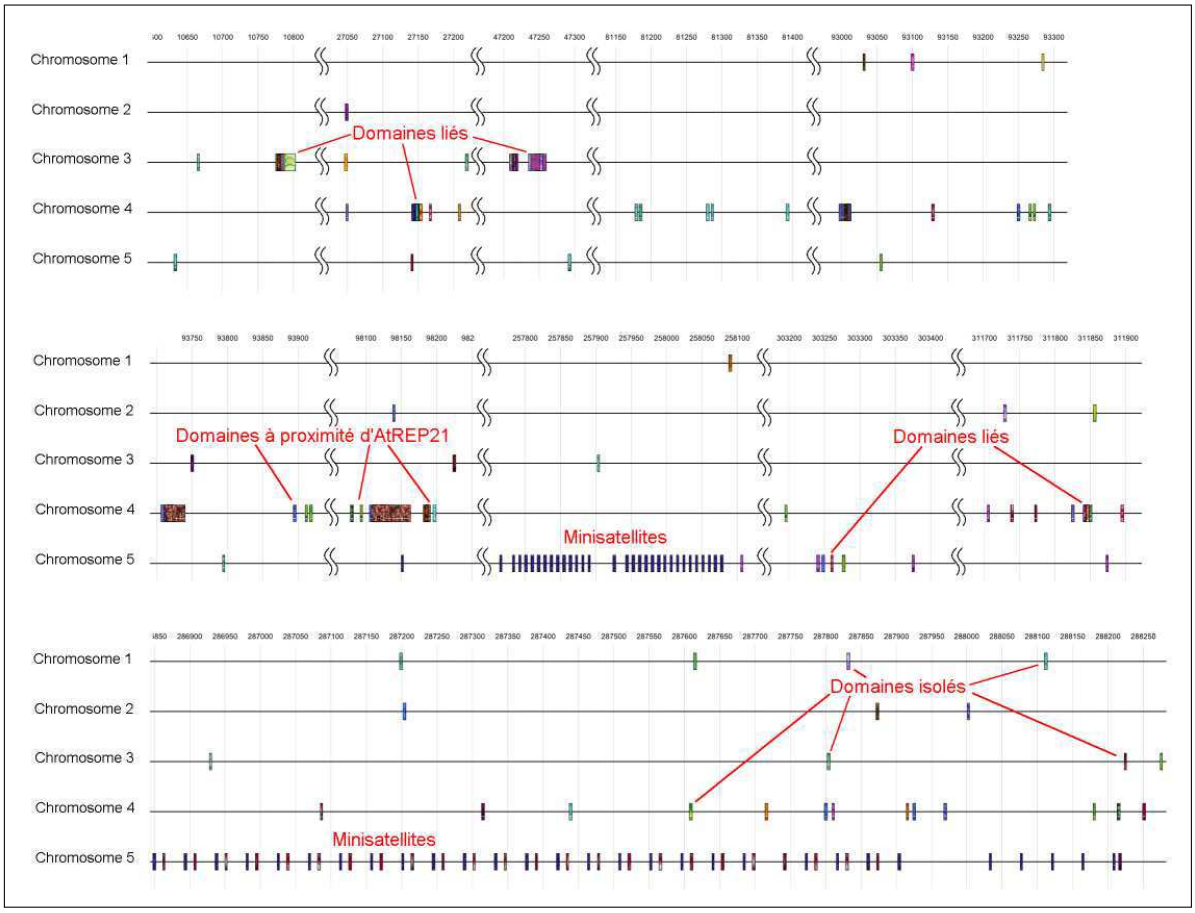


FIG. 3.34 – Exemples d'appariement de domaines à l'extérieur des AtREP21. Une texture particulière a été attribuée à la séquence AtREP21. Les autres textures correspondent aux domaines. L'affichage des domaines internes avant les séquences d'AtREP21 garantit de ne montrer que les domaines externes à AtREP21.

Les domaines 2, 8, 11, 25, 34, 38, 41, 52, 55-i, 57 et 59 sont toujours associés avec

un autre domaine interne d'AtREP21. Les domaines 1, 16, 17, 20, 33, 35, 42, 44, 47, 51 et 55 ont plus de 50 % de leurs occurrences liées à un domaine interne d'AtREP21. Par exemple, le domaine 41 est toujours associé avec le domaine 58. La figure 3.34 montre plusieurs exemples de domaines internes associés dans le génome d'*Arabidopsis thaliana*. Il est intéressant de remarquer qu'à l'exception des domaines 47 et 55-i, aucun autre domaine interne n'est lié de façon majoritaire avec l'une des extrémités hélitroniques (domaines 42 et 44 dans la Figure 3.28, 3.36 et 3.37).

Ces résultats ont montré que la plupart de ces domaines font partie d'une entité biologique plus grande telle qu'un ET. Il est possible qu'un élément transposable (ou un minisatellite) se soit inséré dans une copie d'AtREP21, puisque cette entité a été recopiée avec l'hélitron (Figure 3.35). Au cours du temps, les copies insérées de l'ET (ou du minisatellites) ont divergé. Cette divergence est observée par DomainOrganizer quand il compare deux séquences. Un exemple hypothétique de l'insertion puis de la divergence d'une entité biologique est montré dans la Figure 3.35.

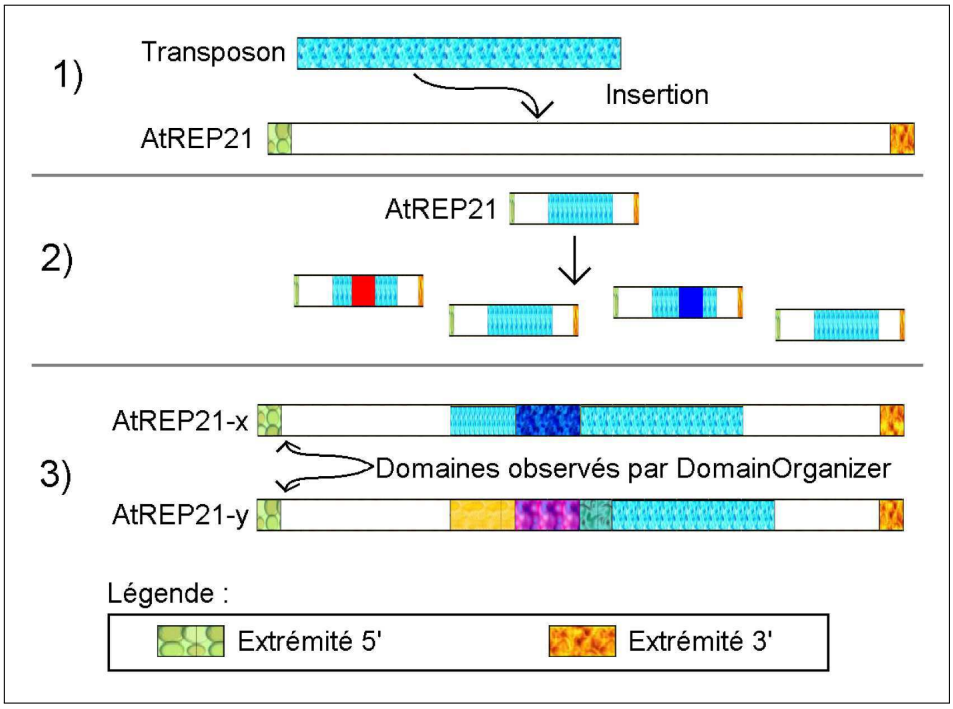


FIG. 3.35 – Exemple hypothétique de création de différents domaines internes dans les hélitrons à partir d'une seule entité biologique insérée. (1) : Un ET s'insère dans une copie d'AtREP21. La copie se multiplie avec l'ET inséré. (2) : La séquence interne d'AtREP21 diverge. (3) : La divergence apparaît avec la comparaison des séquences par DomainOrganizer [196]. L'analyse des domaines internes issus de l'ET montre qu'ils sont associés à l'extérieur de l'hélitron, car ils sont issus de cet ET.

D'autres domaines ne présentent aucune association avec d'autres domaines internes dans le génome d'*Arabidopsis thaliana*. Ainsi, le domaine 58-i est un domaine autonome (possédant son propre mécanisme de copie) qui s'est inséré dans un AtREP21 et qui profite de la transposition de l'AtREP21 pour se multiplier. Le domaine 58-i présente des caractéristiques de minisatellite. Le domaine 9, qui semble avoir été capturé ou résulté de la mutation d'un autre domaine, ne semble pas posséder de méthode de réplication autre que la transposition de l'AtREP21. Les autres occurrences de ce domaine peuvent être

liées à un hélitron tronqué et/ou à un hélitron qui a trop dégénéré pour être identifié.

La figure 3.36 et 3.37 regroupe tous les domaines internes et les informations tirées de ces méthodes. Beaucoup de domaines n'ont pas d'identification par manque de structure et par manque de mode de réplication remarquable. Ces domaines observés par DomainOrganizer peuvent être des artefacts dus à la divergence de la partie de séquence interne des AtREP21. Néanmoins, ces domaines ont permis alors d'observer les "points chauds" (séquence plus soumise à la mutation qu'une séquence normale) de mutations de ces éléments transposables.

Beaucoup de domaines identifiés sont des minisatellites. Ce résultat a déjà été observé dans les hélitrons du riz [86]. Les minisatellites s'insèrent ou sont capturés par les éléments transposables qui leurs permettent ensuite d'envahir les génomes. D'autres domaines identifiés possèdent des structures remarquables telles que les extrémités des hélitrons. Ces identifications ont confirmé la présence de multiples combinaisons de termini dans une même séquence d'hélitron (e.g. section 3.3.3) et ont montré que DomainOrganizer repère les domaines remarquables d'une séquence biologique.

3.5.6 Histoire évolutive de la famille AtREP21

A partir des résultats précédents et d'une étude phylogénétique de la famille AtREP21, nous proposons de retracer les grands événements de l'histoire évolutive de la famille AtREP21.

La phylogénie a été réalisée à partir de l'alignement des séquences obtenu par DomainOrganizer et de la méthode du neighbour-joining [177] (Figure 3.38). Nous avons remarqué une bonne correspondance des groupes d'AtREP21 obtenus avec DomainOrganizer et le regroupement des séquences réalisé par la phylogénie. A l'exception des séquences du groupe 2 mélangées avec les séquences du groupe 1, toutes les autres séquences des autres groupes sont rassemblées dans la phylogénie (Figure 3.38).

L'arbre phylogénétique, tracé par la méthode du neighbour-joining, [177], est en accord avec la classification de DomainOrganizer à l'intérieur d'un même groupe d'AtREP21. Par contre, l'arbre phylogénétique montre une classification totalement différente de celle de DomainOrganizer pour retracer l'histoire évolutive entre les groupes entiers d'AtREP21. Les larges insertions/délétions de domaines internes dans les séquences hélitroniques peuvent fausser la pertinence d'une phylogénie classique.

Les résultats recueillis sur les domaines ont confirmé l'inexactitude de l'arbre phylogénétique. Par exemple, la phylogénie montre que le groupe 7, composé des AtREP21 8 et 27 (Figure 3.38), a évolué à partir du groupe 6 et du groupe 5. L'étude des domaines internes montre au contraire que le groupe composé des AtREP21 8 et 27 ne serait pas issu de ces autres groupes mais serait un groupe ancestral des AtREP21.

Un indice tend à confirmer cette hypothèse : la proportion des domaines internes de ce groupe dans les autres hélitrons. Les domaines 3, 6, 31, 40, 45, 48-50 et 54-ci de ce groupe sont présents dans d'autres hélitrons, alors que la plupart des autres domaines d'AtREP21 ne sont pas présents dans les autres hélitrons (Figure 3.33). Il est plus probable que les domaines 3, 6, 31, 40, 45, 48-50 et 54-ci, qui sont disséminés dans la séquence de ces deux AtREP21, soient des fragments de la séquence ancestrale, plutôt que des échanges successifs de séquences entre l'AtREP21 8 ou 27 (l'un des deux AtREP21 étant probablement une copie de l'autre AtREP21) et un hélitron ancestral.

Ainsi, dans les cas où les différentes copies d'une famille présentent de fortes variations

Domaine	Taille (bp)	% en G+C	Occurrences AtREP21 et chimères	Occurrences Helitron	Occurrences Genome	Lié x fois au domaine y	ΔG / Structure 2nd	Infomation / Identification putative
1	388	24,23	3	3	7	4 / 42	-69,84 / stable	Hélitron3
2	52	30,77	4	4	6	6 / 51-ci	-3,49	
3	36	16,67	2	6	18	2 / 50	-2,09	
4	24	29,17	3	11	170	5 / 7	-0,72	
5	21	33,33	7	41	497	15 / 51-ci	-0,84	Minisatellite
6	26	15,38	3	47	720	33 / 13	0,48	Minisatellite
7	26	30,77	3	5	33	5 / 9	-1,72	
8	36	27,78	4	4	8	8 / 5	-3,27 / hairpin	
9	132	21,97	44	44	51	50 / 59	-14,48 / stable	
10	26	38,46	3	3	15	6 / 9	-3,86 / hairpin	
11	74	21,62	2	2	2	2 / 9	-5,55	Similaire à 5' hélitron
12	25	20,00	5	5	31	10 / 28	-0,31	
13	26	23,07	5	20	191	18 / 6	1,10	Minisatellite
14	26	19,23	3	6	36	2 / 9	-4,05 / hairpin	
15	87	32,18	2	2	13	3 / 13	-10,94 / stable	
16	41	31,71	2	2	3	3 / 6	-4,89 / stable	
17	366	36,61	2	2	3	3 / 15	-59,85 / stable	
18	26	30,77	2	3	21	4 / 17	-7,00 / stable	
19	35	14,29	7	12	74	10 / 12	-2,34	Similaire à 12
20	864	35,88	2	2	4	3 / 18	-145,79 / stable	
21	26	19,23	8	14	145	15 / 25	-1,81	Minisatellite
22	26	19,23	2	4	120	5 / 16	-3,55	
23	37	16,22	3	4	9	3 / 57	-7,29 / hairpin	
24	38	15,79	4	11	24	9 / 9	-6,90 / hairpin	
25	37	16,22	12	12	17	17 / 29	-2,36	
26	26	11,53	7	15	71	12 / 29	-1,01	
27	34	29,41	2	2	5	3 / 53-i	-0,66	
28	25	28,00	12	14	148	17 / 12	0,11	Similaire à 12, Inclus dans 9
29	26	19,23	19	22	68	26 / 51	-2,77	
30	26	11,53	8	34	371	17 / 35	-3,75 / hairpin	Minisatellite
31	26	23,08	2	9	134	3 / 27	-5,34 / hairpin	
32	26	46,15	2	2	8	2 / 31	-3,80	
33	33	24,24	3	3	5	4 / 29	-1,76	
34	25	24,00	14	14	17	17 / 9	-1,67	Inclus dans 9
35	36	19,44	17	17	20	17 / 55	-1,68	Inclus dans 36 / Minisatellite

FIG. 3.36 – Partie 1 du tableau récapitulatif des informations relatives aux domaines internes. L'annotation putative des domaines est donnée si elle a été trouvée.

Domaine	Taille (bp)	% en G+C	Occurrences			Lié x fois au domaine y	ΔG / Structure 2nd	Infomation / Identification putative
			AtREP21 et chimères	Helitron	Genome			
36	455	23,74	1	1	71	1 / 9	-59,22 / stable	
37	25	16,00	3	9	81	6 / 50	-2,05	Minisatellite
38	54	24,07	2	2	2	2 / 40	-5,99	
39	21	38,09	3	7	312	4 / 37	-1,08	Minisatellite
40	25	28,00	2	7	202	10 / 6	-0,10	Minisatellite
41	36	27,77	32	32	34	34 / 58	-2,45 / hairpin	Minisatellite
42	26	53,85	40	46	49	32 / 58	-8,68 / hairpin	Hairpin d'hélitron
43	26	42,31	3	14	19	3 / 41	-3,06	Similaire à 42 / Hairpin d'hélitron
44	99	26,26	41	41	47	46 / 47	-24,89 / stable	5' hélitron
45	25	16,00	5	38	672	20 / 48	-1,24	Minisatellite
46-ci	29	24,14	2	2	32	4 / 27	-2,02 / hairpin	
47	28	32,14	36	36	42	41 / 44	0,09	Chevauche domaine 59
47-ci	20	45,00	10	33	510	16 / 58-i	0,41	Minisatellite
48	25	8,00	10	68	879	33 / 50	-0,02	Minisatellite
49	27	18,52	4	26	350	15 / 60-i	-1,49	
49-i	24	33,33	3	6	182	4 / 56-i	0,14	
50	31	9,68	14	50	533	32 / 48	-3,43 / hairpin	Minisatellite
51	31	22,58	17	17	21	19 / 29	1,14	Similaire à 33
51-ci	23	45,48	7	11	221	11 / 5	1,54	
51-i	26	30,77	1	1	12	1 / 9	-1,40	
52	40	5,00	30	30	35	32 / 9	-4,50 / hairpin	
52-ci	33	15,15	1	1	5	2 / 25	-0,41	
53-i	34	26,47	2	2	7	3 / 27	-3,35	
54-ci	26	23,08	2	4	75	4 / 49	-0,99	
55	32	18,75	16	16	22	18 / 35	-0,88	Inclus dans 36 / Similaire à 35
55-i	31	25,81	37	37	41	41 / 44	-2,60	Chevauche domaine 2
56	32	9,38	32	88	205	21 / 29	-2,40	Minisatellite
56-i	23	39,13	5	10	349	7 / 59-ci	-0,75	
57	33	30,30	30	30	33	33 / 47	0,02	Similaire à 5' hélitron
58	30	6,67	36	56	253	50 / 41	-2,91 / hairpin	Minisatellite
58-ci	29	24,14	5	5	19	8 / 12	-2,93 / hairpin	chevauche domaine 9
58-i	25	12,00	14	63	865	41 / 56	-1,00	Minisatellite
59	41	24,39	32	32	35	35 / 9	-4,32 / hairpin	similaire à 57
59-ci	26	30,77	2	2	23	3 / 16	0,18	
60-i	33	27,27	3	9	15	5 / 49	-1,70	

FIG. 3.37 – Partie 2 du tableau récapitulatif des informations relatives aux domaines internes. L'annotation putative des domaines est donnée si elle a été trouvée.

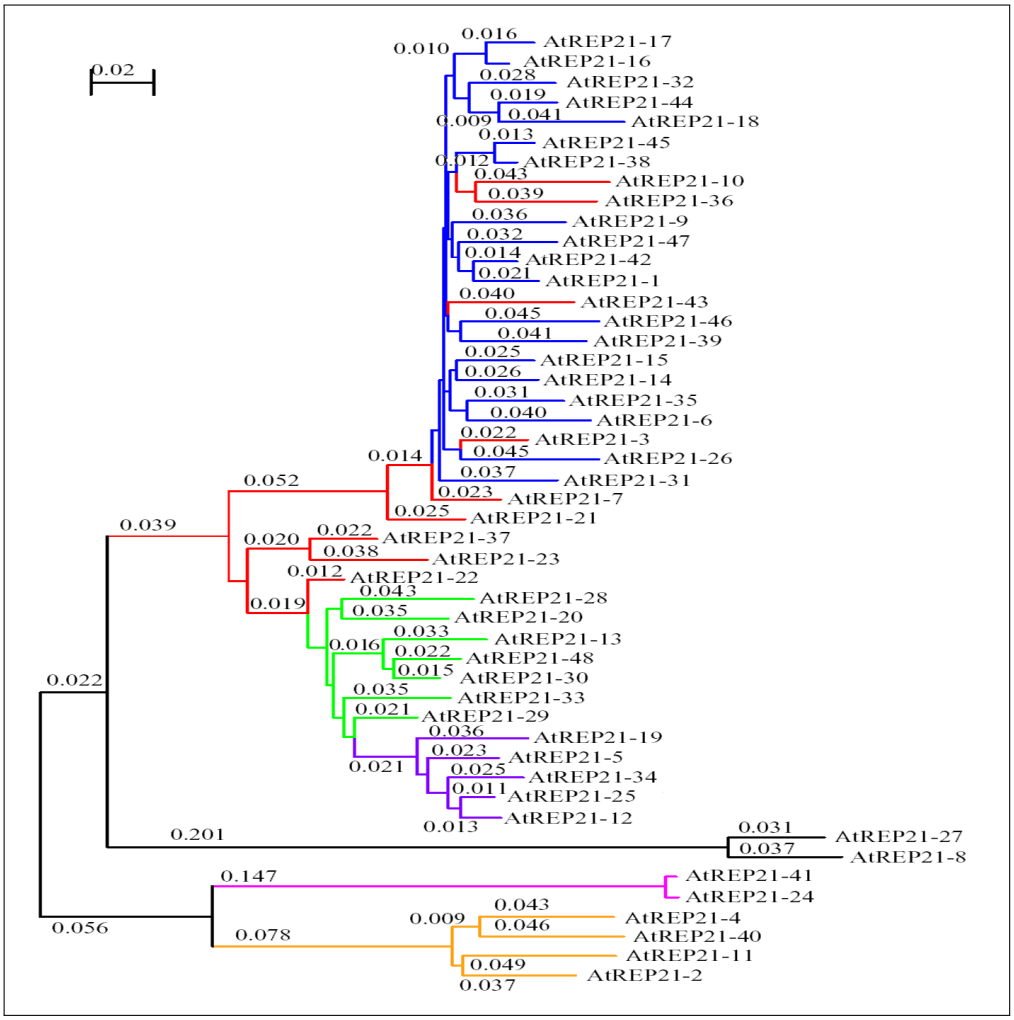


FIG. 3.38 – Arbre phylogénétique des 48 séquences AtREP21 réalisé avec la méthode du neighbour-joining [177]. Les couleurs rajoutées sur l'arbre correspondent aux couleurs des sous-groupes trouvés avec DomainOrganizer (Figure 3.28).

de séquences dues à des insertions ou des délétions de séquences (telles que les éléments transposables ou les familles multigéniques), il est plus pertinent de classer les séquences sur la base de leurs domaines nucléiques que sur la base de leurs séquences entières. Nous pouvons alors suggérer que DomainOrganizer réalise, dans ce cas, un "recodage" compact des séquences qui représente l'identification et l'évolution de chacune des séquences.

A partir de la Figure 3.28, nous proposons le scénario d'évolution suivant, qui nous semble le plus probable par rapport à l'organisation en domaines des séquence d'AtREP21.

- Les AtREP21 8 et 27 (groupe 7) sont apparus à partir d'une autre famille d'hélitrons (autonome ou non). Quelques domaines de la séquence interne auraient été conservés (domaines 3, 6, 31, 40, 45, 48, 49, 50 et 54-ci) et les autres blocs auraient divergé au cours de l'évolution.
- L'autre groupe d'AtREP21 qui semble être un groupe ancestral est le groupe 5 (Figure 3.28). Ce groupe possède aussi des domaines internes répartis dans d'autres hélitrons.
- Il est probable que le domaine 52-ci (séquence palindromique du domaine 52) ait

créé le domaine 26 par dégénérescence.

- Les AtREP21 24 et 41 du groupe 6 sont très proches des AtREP21 du groupe 5 (Figure 3.39). Ces deux groupes présentent une forte homologie de domaines dans les parties 5' et 3' subterminales : plus de 150 pb de domaines communs pour la partie 5' et les domaines 25, 29, 44 et 52-ci pour la partie 3'.
- Le groupe 5 aurait créé ensuite le groupe 4. Ces deux groupes présentent la même combinaison de domaines en 3' (les domaines 29, 38, 41, 42, 52 et 52-ci). Les AtREP21 5 et 19 auraient été les premiers AtREP21 créés de ce groupe, puis seraient apparus les AtREP21 12, 25, 34 qui ont subis la délétion du domaine 29.
- Le groupe 4 a lui-même évolué en séquence du groupe 1 d'AtREP21 (Figure 3.39). Une de ses séquences a subi la délétion du domaine 51, 52-i et 55 et la substitution ou la dégénérescence des domaines 11, 23 et 24 pour le domaine 9.
- Ensuite, le groupe 1 a créé le groupe 2. Le groupe 2 est essentiellement constitué des domaines du groupe 1 avec des délétions et/ou substitutions ponctuelles de domaines (Figure 3.39). Les AtREP21 3, 10, 35 et 43 semblent les premiers AtREP21 de ce groupe avec l'insertion ou la délétion d'un domaine du groupe 1 (Figure 3.28). Puis, la dégénérescence (ou substitution) du domaine 9 aurait créée les copies d'AtREP21 7, 21 et 22. La substitution des domaines 57 et 59 par le domaine 47 et l'insertion des domaines 30 et 35 auraient créés les derniers AtREP21 de ce groupe (AtREP21 23 et 37).
- Enfin, le groupe 3 aurait aussi été créé par le groupe 1, mais il semble que des échanges de domaines aient été réalisés avec le groupe 2.
- Le groupe 2 et 3 sont les deux seuls groupes qui ont subi la substitution des domaines 57 et 59 par le domaine 47 et l'insertion du domaine 35 et/ou 58.

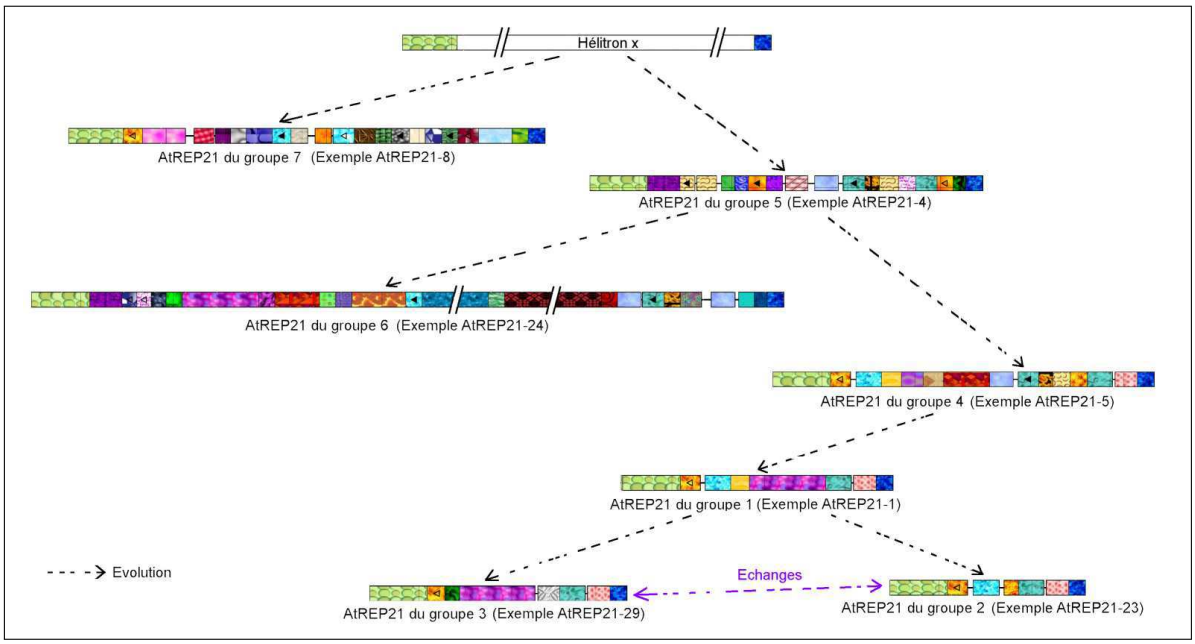


FIG. 3.39 – Evolution de la famille AtREP21 à partir des identifications des domaines internes.

Résumé

Cette étude a montré que les domaines internes des AtREP21 sont peu présents dans les autres hélitrons et semblent donc montrer qu'il y a pas (ou peu) d'échange de domaines entre les familles d'hélitrons. Cette analyse a aussi permis de retracer l'évolution de la famille AtREP21, alors qu'une étude phylogénétique classique n'avait pas pu le faire à cause des grandes insertions/délétions dans la séquence interne. Néanmoins, un grand nombre de domaines internes n'a pas pu être identifié. Cette non-identification de domaines est principalement due à la mauvaise détection des bordures de domaines. Si les séquences à analyser sont très divergentes, les programmes d'alignement multiple (Exemple ClustalW [200]) n'alignent pas correctement les portions de séquences insérées/délétées, en particulier aux niveaux de leurs bordures. Une étude de la structure secondaire des domaines internes a aussi montré une forte corrélation entre la modification de séquence primaire d'un domaine et la structure secondaire d'un AtREP21. Cette corrélation avait déjà été montré dans l'étude de la variation de séquence interne des MITEs *Mariner* [162]. Une recherche systématique de la structure secondaire des domaines détectés pourrait permettre une meilleure définition de ces domaines et conduire à une meilleure analyse de l'évolution des éléments transposables non-autonomes.

Contribution scientifique

Conférence : Tempel S., El Amrani A., Couée I., Nicolas J. 2005. Organisation modulaire des séquences d'ADN répétées. XIII ème Colloque Eléments Transposables. Orsay.

Séminaire : Tempel S., Nicolas J. 15 décembre 2005. Dynamics of Genomes : Modularity in Transposable Elements. ARC : Integrated Biological Networks. Orsay.

Séminaire : Tempel S., El Amrani A., Couée I., Nicolas J. 27 septembre 2005. Organisation en module des éléments répétés. Laboratoire de Parasitologie Moléculaire (UMR 5016 CNRS). Bordeaux.

Poster : Tempel S. Giraud M. Lerman I.C. El Amrani A. Nicolas J. Domain Organization within repeated DNA Application to the study of a new family of transposable elements in Arabidopsis thaliana. 2005. JOBIM.

3.6 Impact sur la régulation de l'expression des gènes du génome hôte

Les études bioinformatiques enlèvent systématiquement les éléments répétés des séquences promotrices ou introniques pour rechercher les motifs de liaison aux facteurs de transcription. La base de la recherche de motifs de régulation *in silico* est la recherche d'un mot "exceptionnel" dans le promoteur : mot répété à l'intérieur des promoteurs de gènes ayant des profils d'expression similaires, mais des séquence promotrices non répétées [175]. Ce contexte de recherche de mots exceptionnels explique la délétion des éléments transposables des promoteurs. Néanmoins, les éléments transposables ont souvent une action sur les gènes à leur proximité [14, 25, 164, 186]. En effet, le travail présenté dans cette thèse avait commencé par la découverte d'un hélitron dans le promoteur du gène ADC1 [62] (chapitre 3.1.1). Il était donc tout naturel de rechercher le rôle des hélitrons sur les gènes à leur proximité.

Dans cette section, nous avons recherché la localisation de tous les hélitrons d'une famille donnée et nous avons sélectionné ceux qui pouvaient avoir une action sur le profil d'expression d'un gène, en particulier les hélitrons présents dans les promoteurs (Figure 3.40). Nous avons ensuite étudié les motifs de régulation présents dans ces promoteurs (hélitrons inclus) et les profils d'expression des gènes associés pour essayer de comprendre l'effet de ces hélitrons sur l'expression génétique (Figure 3.40).

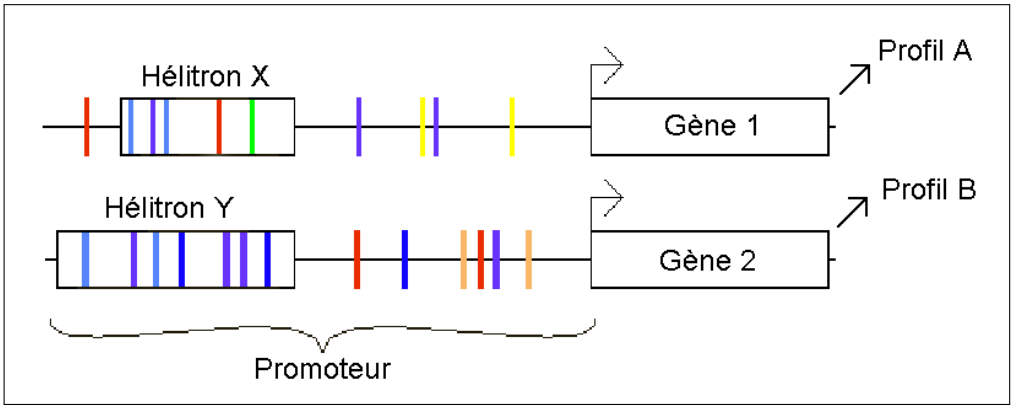


FIG. 3.40 – Modèle d'impact sur la régulation d'un hélitron inséré dans un promoteur de gène. Les différentes barres verticales colorées représentent les différentes boîtes de régulation détectées.

Nous avons limité notre étude aux éléments de la famille AtREP3 (famille dont l'une des copies a été découverte dans le promoteur du gène ADC1 (section 3.1)). Les éléments autonomes à cause de leur taille (supérieure à 10 kb [98]), ont souvent une action délétère sur les gènes à leur proximité. Nous avons donc récupéré, avec STAN [154], toutes les séquences hélitroniques non-autonomes et nous nous sommes limités aux AtREP3 de moins de 3000 pb qui possèdent les deux extrémités hélitroniques d'AtREP3. La grammaire SVG (String Variable Grammar) de cette famille est :

```
TCCTACTATATTATTTGGGAAGTACATTTTAAATGT:9
-x(0,3000)-
AAATCGTCCCGCGGTATACCGCGGGTTAAAATCTAG:9
```

3.6.1 Insertion préférentielle d'AtREP3 dans les promoteurs

Nous avons obtenu 139 séquences AtREP3 complètes dans le génome d'*Arabidopsis thaliana*. Pour chaque hélitron, nous avons recherché le gène en 5' et le gène 3' le plus proche. Pour chaque gène obtenu, nous avons récupéré son nom, sa position et son orientation dans la banque de données TAIR (www.arabidopsis.org). Nous avons classé les AtREP3 non-autonomes en fonction de la nature de leur insertion (promoteur, 3' ou dans le gène) et de leur distance par rapport à un gène (Figure 3.41).

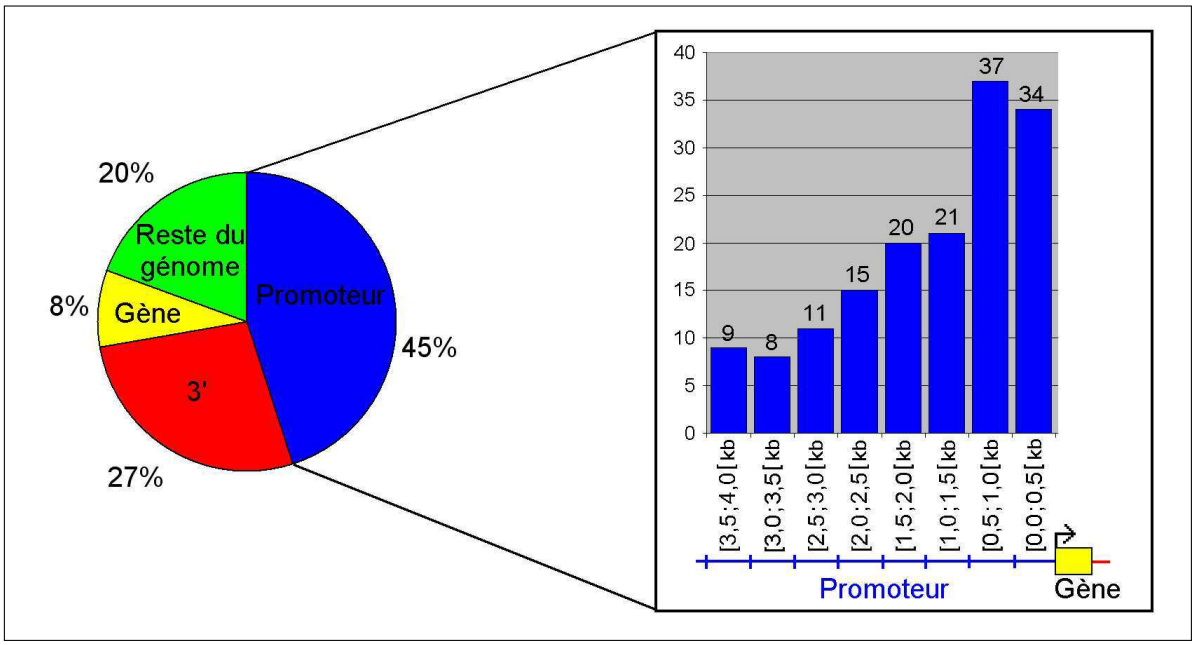


FIG. 3.41 – Répartition des AtREP3 en fonction de la nature de l'insertion et de la distance de l'insertion. Le diagramme à gauche représente les différentes proportions d'insertion des AtREP3 dans les différentes zones du génome. Le graphe à droite donne le nombre d'AtREP3s insérés dans le promoteur par classes de distance au codon START de la transcription.

La Figure 3.41 montre une insertion préférentielle dans le promoteur (45 %). Elle montre également que les AtREP3s (quelles que soient la taille et la composition de l'AtREP3 non-autonome) sont préférentiellement insérés dans le premier millier de paires de bases du promoteur (51 %), et que plus la distance augmente moins AtREP3 est présent dans le promoteur.

Beaucoup d'études montrent que les 3 kb précédant le codon START contiennent la majorité des motifs de liaisons aux facteurs de transcription [36, 139, 193, 155]. L'élément consensus d'AtREP3 mesurant 2 kb [98], nous nous sommes uniquement focalisés sur les AtREP3 insérés à moins de 1 kb des codons START dans les promoteurs des gènes, pour être sûr que l'élément AtREP3 intervient dans son intégralité sur les promoteurs des gènes. Les positions des AtREP3s présents dans cette zone sont répertoriées dans l'annexe 5.4.

3.6.2 (Co-)Régulation tissulaire des gènes à proximité des AtREP3

A partir de la liste de gènes possédant des AtREP3 complets dans leur promoteur, nous avons consulté les bases de données des profils d'expression des gènes d'*Arabidopsis thaliana*. Nous nous sommes limités aux AtREP3 complets.

Même si la séquence interne des AtREP3 est variable, les différentes copies des AtREP3 possèdent des portions de séquences communes ce qui suggèrent que les AtREP3 pourraient avoir un effet régulateur commun. Nous avons recherché l'effet de co-régulateur de ces éléments transposables sur les gènes à leur proximité. Autrement dit, nous avons suggéré que les AtREP3 possédaient des motifs de régulation communs qui induisent une expression similaire des gènes.

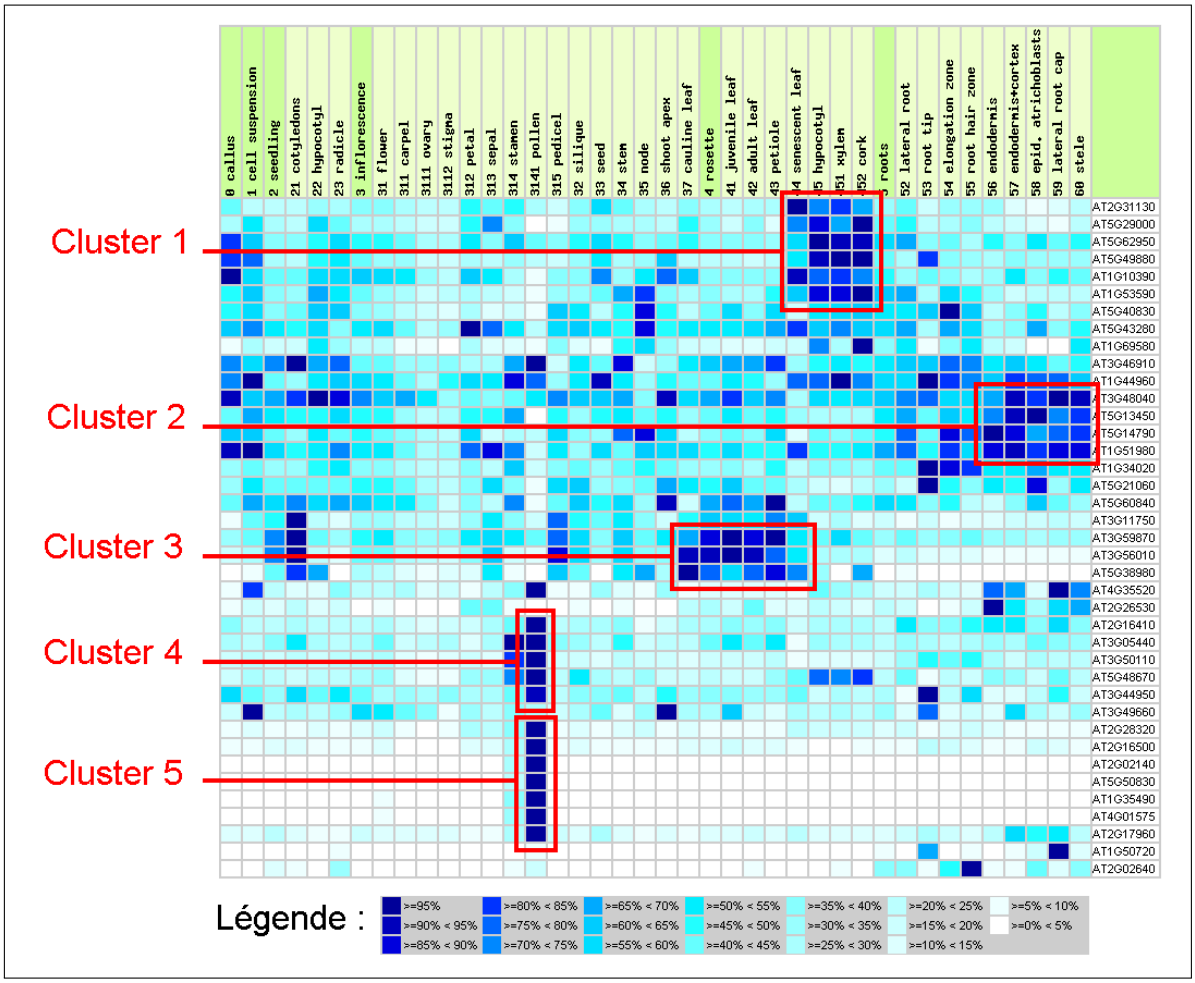


FIG. 3.42 – Profils d'expressions des différents gènes qui contiennent un AtREP3 dans leur promoteur. Les profils observés ont été réalisés sur les différents organes de la plante *Arabidopsis thaliana*. La colorisation des profils de gènes est normalisée. Ainsi, pour chaque gène le plus haut signal d'intensité de transcription obtient la valeur 100 % (couleur bleue-noire) et l'absence de signal obtient la valeur 0 % (couleur blanche).

Nous avons utilisé le site web Genevestigator (www.genevestigator.ethz.ch) qui per-

met d'analyser jusqu'à 500 gènes simultanément dans différentes conditions de stress, à différents stades de croissance et dans différents organes de la plante [212]. Nous avions 81 gènes possédant un AtREP3 à moins de 1 kb du codon START. Néanmoins pour certains gènes, il n'existait pas d'expériences de microarrays à leur sujet. Nous avons donc observé les résultats des profils d'expressions des 47 gènes restants. L'observation des gènes soumis à différents stress et à différents stades de croissance n'a présenté aucun groupement significatif de profils d'expression similaire. Néanmoins, l'observation de la transcription de ces mêmes gènes au niveau tissulaire a montré de grands clusters, très distincts les uns des autres (Figure 3.42). Nous avons choisi d'approfondir l'implication possible des transposons AtREP3 dans cette co-régulation au niveau tissulaire.

La Figure 3.42 montre l'expression tissulaire des gènes à proximité d'AtREP3. Certains gènes ne présentent aucune co-régulation tissulaire. Plusieurs phénomènes biologiques peuvent expliquer ces profils d'expressions uniques malgré la présence d'un AtREP3. La séquence promotrice des gènes étudiés n'est pas uniquement constituée des éléments AtREP3. La partie promotrice non-hélitronique contient des motifs de liaisons aux facteurs de transcription. Ces éléments font varier la régulation des gènes et créent des profils variables d'un gène à l'autre. De plus, les domaines internes spécifiques à certaines copies d'AtREP3 peuvent aussi contenir des motifs de régulation de la transcription.

Ainsi, la plupart des séquences promotrices sont constituées d'une partie promotrice non-hélitronique unique et d'une partie AtREP3 qui comporte aussi des mutations apportant un ensemble variable de motifs de régulation. Cette mosaïque de séquences de régulation explique la grande différence de profils (stress, tissulaire et stade de croissance) entre les différents gènes.

Néanmoins, d'autres groupes de gènes (par exemple le groupe contenant le locus At3G56000) sont regroupés en clusters. Nous avons délimité cinq clusters. Nous nous sommes limités à l'étude du dernier cluster. Ce cluster contient le gène ADC1 (gène à l'origine de la découverte des hélitrons), et présente la particularité que les gènes associés à ce cluster ne sont exprimés que dans le pollen (Figure 3.42). Nous avons nommé ce cluster, le "cluster pollen" par la suite.

3.6.3 Recherche de motifs de régulation dans les promoteurs du cluster pollen

Pour chaque gène du cluster pollen, nous avons cherché à savoir quels motifs de liaison aux facteurs de transcription étaient présents dans le promoteur. Nous avons utilisé l'interface TOUCAN [2] qui nous a permis de rechercher ces motifs selon différentes bases de motifs et avec plusieurs outils de recherche de motif tel que MotifScanner [199] ou de les détecter *ab initio* avec des outils tels que MotifSampler [198].

3.6.3.1 Détection de motifs communs

Nous avons utilisé les banques de motifs PlantCare [175], TFD et AGRIS [45] pour la recherche des motifs de régulation, ainsi que les outils RSATools [204] et MotifSampler [198]. La Figure 3.43 montre les motifs de régulation détectés avec MotifScanner et la base de motifs TFD.

Nous avons sélectionné tous les gènes, présents chez *Arabidopsis thaliana* et dont la transcription ne s'exprime que dans le pollen. Ces gènes représente environ 1 % des gènes

(250 gènes) de la plante et il n'y a que 7 gènes qui contiennent un AtREP3 entier dans le premier millier de paires de base de leur promoteur (Figure 3.42). Nous en déduisons que les AtREP3s n'apportent pas l'élément co-régulateur de l'expression spécifique du pollen, mais que la combinaison des motifs apportés par les AtREP3s et les motifs déjà présents dans le promoteur aboutit à l'expression spécifique dans le pollen.

Nous observons avec toutes les méthodes de recherche, des motifs de liaisons aux facteurs de transcription communs à tous les promoteurs du cluster pollen (7 gènes). Les motifs communs (par exemple ABF1 (ARS-binding factor 1) et Amt (ASA1 mutant), Figure 3.43) sont contenus dans les AtREP3s présents dans les promoteurs du cluster pollen. Aucun motif trouvé uniquement en dehors des AtREP3s n'est commun à tous ces promoteurs. Néanmoins, certains motifs de régulation ne sont pas présents dans tous les AtREP3, mais sont présents dans la partie promotrice non-hélitronique. Aucune de nos recherches de motifs dans ses promoteurs n'a montré de pattern de régulation spécifique au pollen. De plus, tous les motifs détectés dans le cluster pollen sont présents dans les autres promoteurs où AtREP3 est inséré. Nous pouvons suggérer que les AtREP3s n'ont un effet co-régulateur sur le cluster pollen, mais ils peuvent avoir un effet régulateur sur ce cluster.

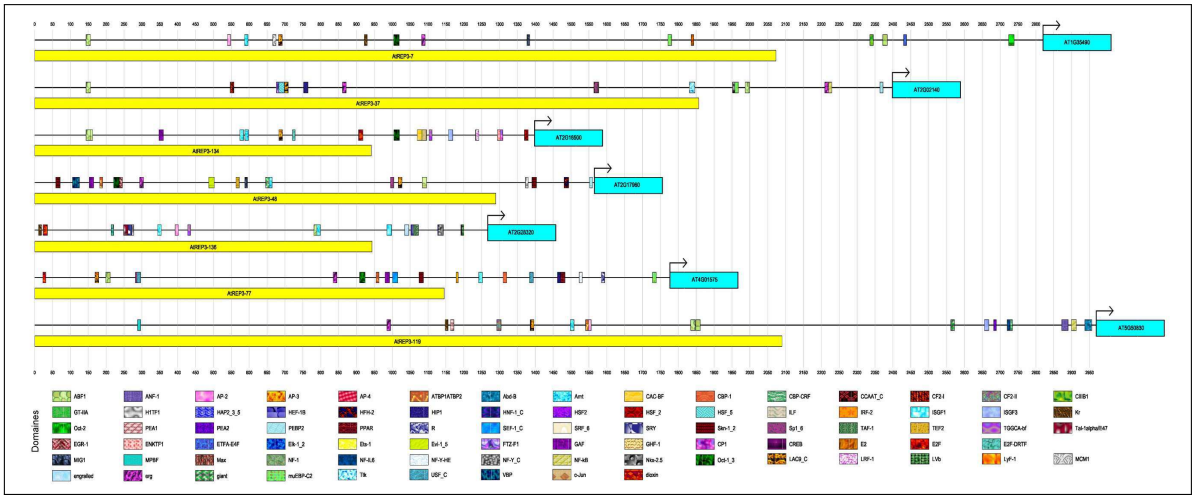


FIG. 3.43 – Motifs de régulation détectés dans les promoteurs des gènes étiquetés par les AtREP3s et ne s'exprimant que dans le pollen de la plante. Les motifs ont été détectés avec l'interface TOUCAN [2] qui a utilisé le logiciel MotifScanner [199] et la base de données TFD. Les emplacements des AtREP3 insérés ont été rajoutés sous les promoteurs. Les gènes présents sur le brin négatif ont été inversés avec leur promoteur pour obtenir la même orientation sur le schéma.

3.6.3.2 Analyse du motif pollen-RE

Une autre étude de l'équipe "Expression génétique et adaptation" avait montré la présence de six motifs pollen-RE (pollen responsive element) dans le promoteur de l'ADC1 (AtREP3 + partie non hélitronique) contre seulement deux motifs pour le promoteur ADC2 [82]. Le nombre plus élevé de pollen-RE dans le promoteur ADC1 était en accord avec l'expression de ce gène dans le pollen, alors que le gène ADC2 s'exprime dans toute

la plante. Néanmoins, la recherche de ce pattern pollen dans d'autres promoteurs n'a pas montré un nombre significativement supérieur de pollen-RE par rapport à ces autres promoteurs. De plus, certains promoteurs, ne s'exprimant pas spécifiquement dans le pollen, intègrent plus le motif pollen que les gènes du cluster pollen. Cet élément de régulation ne semble pas être seul responsable de l'expression spécifique dans le pollen.

3.6.3.3 Absence de motifs de régulation dans le cluster pollen

Nous avons aussi comparé les autres motifs de régulation du cluster pollen avec les autres promoteurs contenant un AtREP3. Nous avons donc vérifié si tous les motifs détectés dans le cluster pollen sont présents dans les autres clusters et inversement. Curieusement, certains motifs de liaison aux facteurs de transcription présents dans les autres promoteurs ne sont pas détectés dans le cluster pollen (Exemple : le motif de régulation E4F). L'absence de ces motifs de régulation dans ce cluster peut expliquer l'absence d'expression des gènes du cluster dans les autres tissus de la plante. Une analyse *in vivo* ultérieure pourrait être réalisée pour vérifier si l'absence de ces motifs est bien responsable de l'absence d'expression dans les autres tissus.

3.6.3.4 Effets pollen-spécifique sans présence d'AtREP3

Pour vérifier si ce cluster pollen était particulier, nous avons commencé à analyser l'ensemble des gènes d'*Arabidopsis thaliana* et recherché d'autres gènes ne s'exprimant que dans le pollen.

Environ 1 % des gènes ne s'expriment que dans le pollen. Néanmoins, ces gènes ne possèdent pas tous des AtREP3s insérés dans leur promoteur. Ce résultat a montré que les AtREP3s (avec les deux termini) ne sont pas les seuls responsables de ce profil d'expression particulier.

Résumé

Les AtREP3s, par extension les hélitrons, sont majoritairement insérés dans les promoteurs des gènes. Ils contiennent dans leur séquence des motifs de liaison aux facteurs de transcription (Figure 3.43). A l'exception de l'AtREP3 présent dans le gène ADC1, aucun résultat n'a pu montré leur rôle régulateur ou co-régulateur des AtREP3s. Si la démonstration de leur rôle co-régulateur semble compromis, leur rôle régulateur est toujours envisageable. Les résultats précédents semblent montrer deux voies possibles d'actions régulatrices des AtREP3s sur les gènes : une voie combinatoire et une voie délétère. La voie combinatoire correspond à la modification du profil d'expression du gène par la combinaison des motifs de régulation existants dans le promoteur avant l'insertion de l'ET et des motifs apportés par l'AtREP3. La voie délétère (non antagoniste avec la précédente) correspond à la perte de motifs de régulation par l'insertion de l'ET : l'insertion éloigne une partie du promoteur proximal (proche du gène) du codon START et peut rendre inaccessible les motifs présents dans cette partie du promoteur.

Si nous voulons savoir si les hélitrons agissent sur la transcription des gènes à leur proximité, de nouvelles analyses devront être effectuées, aussi bien au niveau bioinformatique que biologique. Certaines de ces expériences sont évoquées dans le chapitre suivant.

Conclusion et Perspectives

4.1 Conclusion

Tous les précédents résultats ont montré l'importance d'une modélisation syntaxique des éléments transposables, basée sur la description des extrémités. Nous avons focalisé notre travail sur l'analyse des hélitrons, qui sont des ETs très difficiles à identifier avec des méthodes d'alignement de séquences. Cette modélisation a débouché sur la détection exhaustive de tous les hélitrons dans le génome d'*Arabidopsis thaliana* et la création d'une nouvelle nomenclature des hélitrons. Cette nomenclature est plus adaptée aux interactions biologiques autonomes non-autonomes que la précédente classification des hélitrons [98].

Un point essentiel de cette thèse a porté sur la définition et la détection d'un concept de domaine nucléique. La création de l'outil DomainOrganizer [196] automatise l'identification de ces domaines dans une famille de séquences répétées. L'évolution d'une famille d'éléments transposables non-autonomes a été pour la première fois retracée grâce à l'analyse systématique des domaines internes de cette famille. De plus, la courte analyse d'un MITE montre que ce concept de domaines nucléiques peut être étendu à l'ensemble des éléments transposables non-autonomes de classe II. Il serait intéressant de tester aussi DomainOrganizer avec des éléments de classe I, tels que les SINEs ou les LINEs qui possèdent des domaines nucléiques connus.

Ensuite, l'analyse de la localisation des hélitrons chez *Arabidopsis thaliana* a montré une insertion préférentielle de ceux-ci dans les promoteurs des gènes. De plus, toutes les analyses de recherche de motifs de régulation montrent que certains motifs sont contenus dans les hélitrons non-autonomes. Néanmoins, l'analyse des profils d'expression des gènes, qui contiennent ces hélitrons non-autonomes, ne montre pas de façon évidente une influence régulatrice systématique des hélitrons. Des études complémentaires *in vitro* doivent être réalisées pour comprendre le rôle des hélitrons insérés dans les promoteurs de gènes.

A partir des nombreux résultats précédents, de nouvelles interrogations se posent. Est-ce que les hélitrons ont une action régulatrice sur les gènes à leur proximité ? La protéine RPA et l'hélicase sont-elles suffisantes pour la transposition des hélitrons et comment fonctionne réellement la transposition des hélitrons ? Est-ce que les hélitrons non-autonomes sont les seuls éléments transposables à évoluer par domaines nucléiques ou est-ce que le génome entier évolue aussi par modules ? Quels sont les motifs de liaison aux facteurs de transcription qui déterminent le profil d'expression du gène ?

Dans les deux premières études (la nomenclature des hélitrons et la découpe des hélitrons en domaines nucléiques), l'optimisation combinatoire des données a permis de sélectionner, parmi un large éventail de possibilités, le jeu de données minimal qui per-

met de représenter l'ensemble des données traitées. Nous proposons alors de déterminer le jeu minimal de motifs de régulation nécessaires pour obtenir l'ensemble des profils d'expression des gènes, en appliquant également une formalisation et des techniques issues de l'optimisation combinatoire.

Les deux approches, biologiques et bioinformatiques, sont nécessaires pour répondre à l'ensemble de ces questions.

4.2 Perspectives biologiques

L'approche biologique permet de confronter les hypothèses émises par les méthodes bioinformatiques aux mécanismes biologiques qui restent largement à élucider. A partir de cette thèse, deux grandes études seraient à réaliser : le mécanisme de transposition des hélitrons et le rôle des hélitrons dans le génome d'*Arabidopsis thaliana*.

4.2.1 Modèle d'étude de la transposition des hélitrons

Au début de la thèse, seule la présence des hélitrons dans les génomes était connue et leur mode de transposition était hypothétique [98, 70]. Actuellement, aucun article n'a démontré le fonctionnement de la transposition des hélitrons. La seule découverte publiée sur le mode de transposition des hélitrons est qu'ils transposent encore dans certains génomes de plantes [119, 32].

Parmi les différents objectifs possibles, deux aspects du mécanisme de transposition des hélitrons nous semblent particulièrement importants à étudier : le rôle de chaque protéine au moment de la transposition et la reconnaissance des extrémités hélitroniques par les protéines de transposition.

4.2.1.1 Rôle des protéines de transposition des hélitrons autonomes

La découverte d'un ORF inconnu chez certains hélitrons (Figure 3.13) dans le génome d'Arabidopsis et la découverte d'un ORF non décrit précédemment chez la chauve-souris [169] semblent montrer que toutes les protéines nécessaires à la transposition des hélitrons ne sont pas décrites. Il serait donc particulièrement intéressant d'analyser chacun des ORFs détectés dans les hélitrons, et de comprendre leur(s) rôle(s) lors de la transposition.

Nous proposons une expérimentation basée sur la production de cellules transformées d'Arabidopsis contenant deux constructions introduites par la bactérie *Agrobacterium*. Chacune de ces constructions contiendrait un gène de résistance à un antibiotique pour sélectionner les cellules qui ont intégré les deux constructions (Figure 4.1). Sur une construction (que nous nommerons "construction-transposable"), on incorporerait un gène de résistance dans lequel serait inséré un hélitron non-autonome. Sur l'autre construction (nommé "construction-ORF"), on incorporerait les gènes du complexe de transposition d'un hélitron mais on retirerait ses extrémités hélitroniques pour empêcher cette séquence d'être transposée et on incorporerait un autre gène de résistance (Figure 4.1). Chaque cellule serait séparée par cytométrie de flux et se développerait sur un milieu sélectif.

Avec ce montage, nous proposerions de réaliser un stress sur les cellules transformées, à cause la transposition de l'hélitron non-autonome, le gène de résistance de la construction transposable serait de nouveau actif. Les cellules, où la transposition aurait eu lieu, survivraient aux toxines du milieu de culture. Il suffirait alors de compter les souches de

cellules qui se seraient développées pour estimer la taux de transposition d'un hélitron non-autonome dans une cellule stressée d'*Arabidopsis thaliana*. Si l'on faisait varier les ORFs de l'hélitron sur la construction-ORF, nous vérifierions les ORFs qui sont indispensables à la transposition d'un hélitron et le rôle de chacun d'eux. La délétion choisie d'une extrémité pourrait aussi confirmer les rôles des deux extrémités dans la transposition des hélitrons.

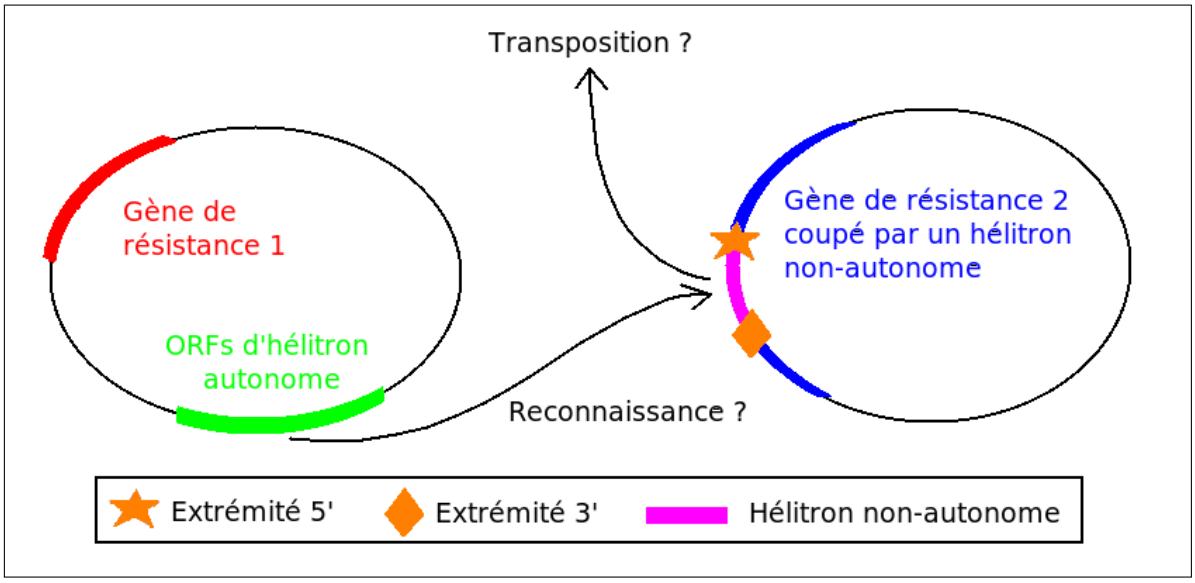


FIG. 4.1 – Observation de l'activité de transposition par l'intégration de deux constructions transformées dans une cellule de plante. La construction de droite contient un gène de resistance dans lequel est inséré les extrémités hélitroniques pour la reconnaissance des protéines de transposition. La construction de gauche contient un hélitron autonome dépourvu de ses extrémités pour éviter sa propre transposition. La construction de droite contient aussi un gène de résistance. Seules les plantes transformées avec les deux constructions survivront sur un milieu de culture sélectif. Nous pourrons modifier la séquence de l'hélitron autonome pour enlever un ORF donné et observer l'importance de cet ORF sur l'efficacité de la transposition.

4.2.1.2 Choix de la combinaison de termini lors de la transposition

Nous avons découvert qu'un hélitron contient plusieurs combinaisons d'extrémités hélitroniques (Figure 3.13). Nous avons choisi l'ensemble de couples de termini minimum qui couvre l'ensemble des séquences des hélitrons et avons créé une nouvelle nomenclature (paragraphe 3.3.4). Néanmoins cet ensemble n'est peut être pas l'ensemble de termini utilisés par les protéines de transposition. Nous proposons donc un protocole qui permet de découvrir quelle extrémité est reconnue par les protéines de transposition.

A partir des hélitrons autonomes capables de transposer (après une vérification expérimentale), nous créerions un hélitron non-autonome possédant toutes les extrémités 5' et 3' des hélitrons autonomes et créerions des hélitrons autonomes sans extrémité. Nous réutiliserions le système précédent à double construction, l'un contenant l'hélitron non-autonome et l'autre contenant un des hélitrons autonomes. Cette construction serait

différente en fonction du complexe de transposition que l'on voudrait étudier. Nous observerions ensuite pour un hélitron autonome donné (pour une combinaison de termini donnée) le taux de transposition de l'hélitron non-autonome en fonction de la combinaison de termini qui transpose (Figure 4.2). Les différents taux de transposition mesurés permettraient de savoir si les protéines de transposition reconnaissent une séquence particulière de termini ou reconnaissent toutes les extrémités hélitroniques mais avec une association préférentielle pour certaines d'entres elles (Figure 4.2).

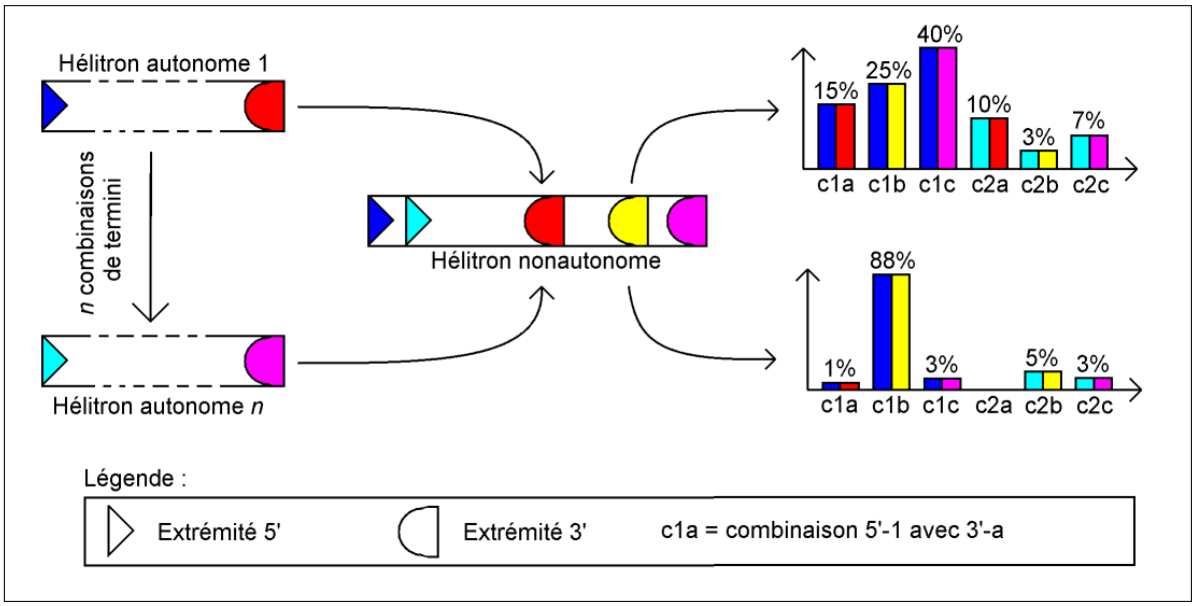


FIG. 4.2 – Analyse de la préférence des extrémités hélitroniques lors de la transposition. Il s'agit d'observer le taux de transposition de chaque combinaison de termini présente dans l'hélitron non-autonome. Les graphiques à droite de la figure montrent les deux cas possibles de taux que l'on pourrait obtenir : le graphique du haut montre une reconnaissance préférentielle et le graphique du bas présente une reconnaissance spécifique des extrémités.

4.2.2 Modèle de confirmation de l'action (co-)régulatrice des hélitrons

L'AtREP3 a une action sur la régulation du gène ADC1 [62, 82]. Néanmoins, son rôle exact n'est pas connu. Nous proposons deux types d'association des AtREP3s avec un gène marqueur (Exemple : gène GUS) : la première association serait de créer un promoteur composé d'AtREP3 et d'un promoteur minimal (Figure 4.3), la deuxième association serait de copier le promoteur d'un gène contenant un AtREP3 (Exemple ADC1). Comme il serait trop long d'associer les 141 AtREP3s détectés avec le gène rapporteur, nous proposons de limiter le nombre d'AtREP3 au nombre de formes différentes d'AtREP3 : un AtREP3 entier mesurant environ 2100 pb (similaire au consensus de Repbase [98]), un AtREP3 tronqué de sa partie 3' (par exemple l'AtREP3 de l'ADC1) et un AtREP3 tronqué de sa partie 5'. Ces trois formes existent pour le cluster pollen.

Pour ces deux types d'association, l'analyse du gène marqueur permettrait de connaître

l'action régulatrice des AtREP3s sous ses trois principales formes. L'association AtREP3, promoteur minimal et gène marqueur permettrait de savoir si l'AtREP3 est suffisant pour induire l'activité transcriptionnelle d'un gène. Pour l'association promoteur entier et gène rapporteur, nous réaliserions des mutations par délétion dans les différentes parties du promoteur pour connaître le rôle transcriptionnel de chaque portion du promoteur. Cette analyse permettrait de savoir s'il existe une relation entre les différentes régulations des gènes par les hélitrons et la variabilité en domaines des hélitrons.

FIG. 4.3 – Modèle d'étude de l'activité régulatrice de l'AtREP3 présent dans le promoteur de l'ADC1. L'AtREP3, avec la séquence promotrice, a un rôle régulateur sur le gène ADC1 [82, 62] (schéma de gauche). L'AtREP3 de l'ADC1 est inséré en amont d'un promoteur minimal et d'un gène marqueur. Nous pouvons ensuite mesurer l'action régulatrice de cet hélitron, indépendamment des zones non-hélitroniques du promoteur ADC1.

4.3 Perspectives bioinformatiques

L'approche bioinformatique permet une approche globale (à l'échelle d'un génome entier) des questions posées. Quatre grands axes de recherche bioinformatique peuvent continuer les travaux de cette thèse : préciser le rôle des hélitrons sur la régulation des gènes d'*Arabidopsis thaliana*, créer un modèle syntaxique général pour les hélitrons, mettre au point une méthode d'analyse de la dynamique d'évolution par domaines des éléments transposables et/ou d'un génome entier et optimiser la segmentation des séquences en domaines.

4.3.1 Analyse à l'échelle du génome entier de la (co-)régulation des gènes par les hélitrons

Les derniers résultats de cette thèse sur le rôle régulateur des hélitrons sont assez contradictoires. Certains résultats bioinformatiques (par exemple la présence de motifs de liaison aux facteurs de transcription communs dans les gènes du cluster pollen 3.6.3.1) indiquent que les hélitrons apportent un rôle co-régulateur aux gènes. Certains résultats biologiques montre une corrélation de l'action régulatrice des AtREP3s sur les gènes [82]. Néanmoins, d'autres résultats n'expliquent pas ou contredisent la régulation des gènes par les AtREP3s (par extension pour les hélitrons). Ainsi, les gènes du cluster pollen ne semblent pas devoir leur profil expression aux AtREP3s. Ces mêmes gènes n'ont aucune co-régulation si nous observons les données de puces à ADN sur les expériences de stress et de la croissance de la plante. De plus, beaucoup d'autres gènes ayant la même expression tissulaire que ce cluster ne semblent pas présenter d'AtREP3.

L'influence des hélitrons sur les gènes peut être vérifiée sur les clusters tels que le cluster pollen. S'il est vrai que d'autres gènes possèdent le même profil d'expression tissulaire que ce cluster, mais ne possèdent pas d'AtREP3 entier, nous n'avons pas testé la présence d'AtREP3 tronqués (voire même d'autres hélitrons) ou la présence de séquence interne appartenant à la famille AtREP3 (Figure 4.4).

Dans un premier temps, il faudrait sélectionner l'ensemble des promoteurs des gènes ayant un profil d'expression tissulaire limité au pollen, puis comparer les séquences les unes aux autres pour détecter un (ou des) motif(s) commun(s). Il nous faudrait aussi comparer chacun des promoteurs aux séquences des AtREP3s (ou des autres hélitrons) (Figure 4.4).

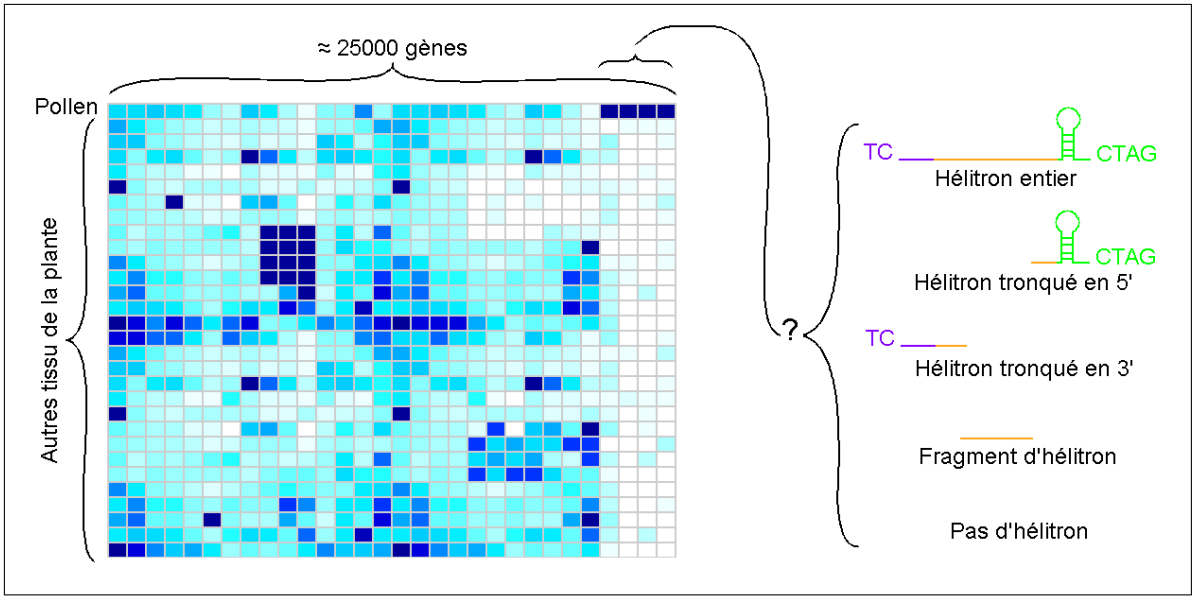


FIG. 4.4 – Recherche d'un hélitron dans l'ensemble des gènes spécifiquement transcrits dans le tissu pollen. Le résultat de cette puce à ADN hypothétique représente schématiquement l'ensemble des profils tissulaires des gènes d'*Arabidopsis thaliana*. Une ligne représente un tissu, une colonne représente un gène. Pour tous les gènes exprimés uniquement dans le pollen (en haut à droite de la puce), une recherche systématique d'hélitron ou de fragment d'hélitron devra être réalisée.

Cette analyse permettrait de connaître le nombre réel de gènes qui présentent un hélitron dans leur promoteur et qui ne s'expriment que dans le pollen. La proportion de gènes ayant un hélitron permettrait de mesurer l'influence de ces hélitrons sur ce profil d'expression. Des études similaires seraient à effectuer sur les autres clusters tissulaires trouvés (Figure 3.42) et devraient être validées avec les expérimentations biologiques (paragraphe 4.2.2).

4.3.2 Recherche d'un modèle syntaxique général pour les hélitrons

La modélisation des hélitrons chez *Arabidopsis thaliana* a montré que les extrémités étaient suffisantes pour détecter les hélitrons. Les modèles des différentes familles sont

très variables, mais ils présentent aussi quelques points communs. Ces motifs communs entre familles pourraient aider à concevoir un modèle syntaxique unique pour les hélitrons présents chez Arabidopsis, mais aussi pour ceux présents dans les autres génomes.

Ainsi, lors de la conception du modèle hélitronique, nous avons voulu intégrer sans succès la structure secondaire de l'hairpin (paragraphe 3.2.3). L'analyseur STAN [154] détectait beaucoup trop de faux positifs à cause de la recherche de n'importe quel hairpin. En fait, nous pensons que la recherche de cette structure secondaire reste impérative mais devrait être couplée avec la séquence primaire riche en G+C. Malheureusement, aucun analyseur actuel, à notre connaissance, ne permet de rechercher en même temps une structure secondaire ou tertiaire qui contienne une séquence primaire particulière.

L'équipe Symbiose met au point un nouvel outil syntaxique (nommé LOGOL) qui permettrait entre autres de rechercher simultanément une séquence et une structure et qui devrait permettre d'aborder de telles questions.

4.3.3 Optimisation de la segmentation en domaines des séquences répétées

L'analyse de la segmentation des domaines de la famille AtREP21 a montré une découpe imparfaite des domaines tant au niveau du nombre de domaines qu'au niveau de la limite de chaque domaine.

Nous avons comparé notre méthode de segmentation avec une autre méthode réalisée par l'équipe Symbiose : Pygram [57]. Cette méthode détecte les répétitions maximales exactes contenues à l'intérieur d'une séquence, et par rapport aux autres séquences. La Figure 4.5 montre la comparaison des deux méthodes pour trois AtREP21. Pour Pygram, nous avons choisi de ne visualiser que les répétitions exactes de taille supérieur à 16 nucléotides.

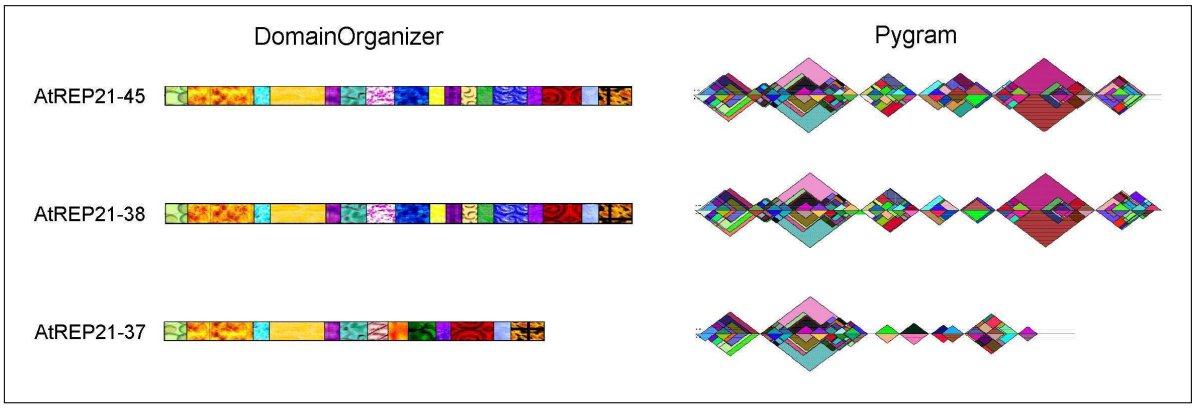


FIG. 4.5 – Comparaison de la segmentation obtenue par DomainOrganizer et Pygram pour trois AtREP21. DomainOrganizer représente chaque domaine par une texture différente. Deux séquences son considérés appartenant au même domaine s'il sont similaires à plus 80 %. Pygram visualise les répétitions exactes avec des triangles, une couleur de triangle est attribué à chaque domaine. La hauteur des triangles est proportionnel à la taille des séquences.

Nous observons que les deux premiers AtREP21 sont similaires avec la méthode DomainOrganizer et Pygram : DomainOrganizer montre la même combinaison de domaines

et Pygram montre les mêmes pyramides pour les deux séquences. Des différences notables entre les deux méthodes apparaissent pour l'AtREP21 37. Cette différence provient de la nature des domaines recherchés : exacts pour Pygram et approchés avec DomainOrganizer. Pygram a l'avantage de ne pas être tributaire de la taille des domaines puisqu'il recherche toutes les répétitions quelle que soit leur taille. De plus, Pygram délimite parfaitement les répétitions, car il utilise un index à la place d'un alignement multiple.

Nous proposons de remplacer l'alignement multiple par une découverte directe des domaines par Pygram. A partir de la liste des domaines exacts, nous proposerions de créer un algorithme de fusion automatisée des répétitions qui permettra d'obtenir des domaines approchés. Cette algorithme pourrait tenir compte de la structure secondaire des domaines pour assembler les différents domaines. L'étude des structures secondaires des domaines d'AtREP21 a montré une corrélation entre la pertinence biologique d'un domaine et sa structure. Cette approche (Pygram + Structure secondaire) permettrait de s'affranchir des paramètres passés en arguments par l'utilisateur et d'obtenir un seul résultat optimisé pour un jeu de séquences donné.

4.3.4 Analyse des domaines nucléiques à l'échelle d'un génome entier

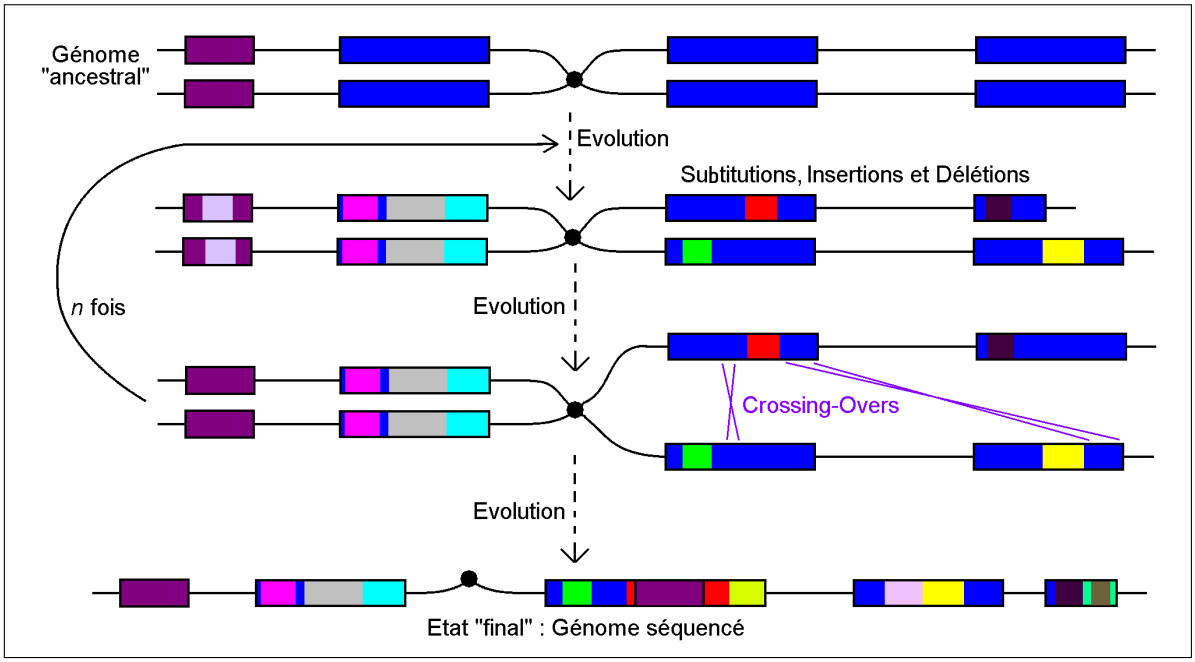


FIG. 4.6 – Evolution schématique des éléments transposables d'un génome eucaryote. A partir du génome ancestral, les différentes copies d'éléments transposables subissent des mutations comme les autres séquences d'ADN, mais subissent aussi des recombinaisons "illégitimes" dues à leur nature de séquences répétées. Le nombre de ces mutations et recombinaisons est variable d'une espèce à l'autre. Après le séquençage d'un génome, on observe une mosaïque d'éléments distincts dont il est très difficile de reconnaître que ces éléments sont issus de la même famille d'ET.

L'analyse des copies de la famille AtREP21 par DomainOrganizer [196] et l'analyse

manuelle de la famille AtREP3 ont montré que l'évolution de ces deux familles est principalement dues à des substitutions, des insertions ou des délétions de blocs de nucléotides que nous avons appelés domaines. Ces précédents résultats ont mis en évidence l'évolution en domaines des hélitrons non-autonomes. Les autres éléments transposables non-autonomes semblent aussi évoluer par blocs de domaines (Figure 3.27). Si nous pouvons connaître les grandes lignes évolutives d'une famille d'élément transposable, il est très difficile de découvrir l'évolution d'un type d'éléments transposable comme les hélitrons ou les *Mariner*. Il est donc encore plus difficile de connaître l'histoire de toutes les familles d'éléments transposables dans un génome eucaryote. Les éléments transposables jouant un rôle essentiel dans la structure du génome (Introduction de la thèse), cette connaissance est nécessaire pour appréhender la dynamique évolutive des génomes eucaryotes.

Pour connaître la dynamique des ETs à l'échelle du génome, la première étape est d'identifier tous les domaines nucléiques appartenant (ou ayant appartenu) à des éléments transposables. Cette étape est très complexe à réaliser, car les ETs ont subi de nombreuses mutations et recombinaisons à chaque nouvelle duplication du génome (Figure 4.6).

Après l'identification de tous les domaines, nous pensons qu'il est possible de travailler sur un algorithme de "reverse engineering biologique" qui permettrait, à partir du génome séquencé, de retracer l'histoire évolutive et de retrouver la répartition des éléments transposables dans le génome ancestral. Cette histoire évolutive serait un bon indicateur de l'histoire structurelle du génome eucaryote, due à l'importance des éléments transposables dans la dynamique évolutive d'un génome.

5.1 Annexe 1 : Positions des Hélitrons dans le génome d'*Arabidopsis thaliana*

Le tableau donne pour toutes les positions des combinaisons des Hélitrons détectées. Ainsi une même position peut avoir plusieurs combinaisons possibles et les différentes combinaisons peuvent aussi se chevaucher. Les extrémités 5' correspondant aux nombres et les extrémités 3' correspondent à des lettres. Le tableau se lit de gauche à droite et de haut en bas.

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
1	17009	18729	14_u	1	499373	501371	14_a,e,f,g,o,i,l,m
1	540433	540986	5,6,16_o,i,l,m	1	696474	697373	9,5,6_a,e,g,t,o,i,l,m
1	1160223	1161224	1,9,5,6_a,c,e,f,g,k,t,o,i,l,m	1	1275792	1276611	17_p,s
1	1705400	1706437	9,5,6_e,f,g,o,i,l,m	1	2233819	2235165	14_u
1	2587210	2587829	9,5,6_e,o,i,l,m	1	2587210	2588425	5,6_e,o,i,l,m
1	2587829	2588425	5,6_a,e,g,t,o,i,l,m	1	3023485	3024294	17_p
1	3342447	3343054	17_s	1	3413554	3414840	14_a,e,f,g,o,i,l,m
1	3497753	3498865	11_a,e,f,g,h,t,v,o,i,l,m	1	3497753	3498957	5,6_a,e,f,g,h,t,v,o,i,l,m
1	4770112	4770661	16,5,6_e,o,i,l,m	1	4785458	4787251	14_o,i,l,m
1	5436907	5437782	9,5,6_e,t,o,i,l,m	1	5531290	5531885	5,6_e,o,i,l,m
1	5657937	5659261	18,19_q	1	5862942	5863504	1,5,6_e,o,i,l,m
1	5899199	5899747	16,5,6_e,o,i,l,m	1	6074878	6075487	17_k,s
1	6233929	6234712	18,19_q	1	6233929	6235091	18,19_q
1	6402912	6403648	9,5,6_t	1	6403869	6405383	17_p
1	6735855	6737355	14_u	1	7035181	7037279	14_a,e,f,g,h,t,v,o,i,l,m
1	7281711	7284070	1,16_a,e,f,g,v,o,i,l,m	1	7833173	7834131	17_p,s
1	8114759	8115698	14_o,i,l,m	1	8169443	8170339	17_p
1	8189950	8190547	16,5,6_e,t,o,i,l,m	1	8235855	8236988	17_p
1	8289493	8291559	14_u	1	8374541	8375139	22_h,v,o,i,l,m
1	8473358	8490182	14_q	1	8489403	8490182	14_a,e,f,g,h,t,v,o,i,l,m
1	8538878	8539751	5,6_e,f,g,t,o,i,l,m	1	8577157	8578054	17_p
1	8622745	8623596	9,5,6_e,o,i,l,m	1	9062071	9062506	22_v,o,i,l,m
1	9135743	9136642	17_p	1	9135743	9138304	17_p
1	9136642	9138304	17_p	1	9384147	9385019	5,6_e,t,o,i,l,m
1	9418050	9418932	5,6_e,o,i,l,m	1	9547128	9551043	12_j
1	9816197	9817137	14_o,i,l,m	1	9816197	9818260	14_a,e,f,g,h,t,v,o,i,l,m
1	9836495	9837567	9,5,6_a,e,g,k,o,i,l,m	1	9862682	9863288	17_s
1	9865331	9867400	14_a,f,g,o,i,l,m	1	9866455	9867400	14_o,i,l,m
1	9918440	9919318	16,5,6_f,o,i,l,m	1	10021400	10021947	16,5,6_e,f,h,v,o,i,l,m
1	10160979	10161820	5,6_o,i,l,m	1	10221825	10222357	5,6_h,k,t,o,i,l,m
1	10221825	10240181	5,6_e,g,o,i,l,m	1	10239306	10240181	9,5,6_e,g,o,i,l,m
1	10524701	10529652	12_j	1	10568159	10580124	17_a,e,f,g,h,t,v,o,i,l,m
1	10579408	10580124	9,5,6_a,e,f,g,h,t,v,o,i,l,m	1	10579408	10598298	9,5,6_k,s
1	10591412	10591965	5,6_e,o,i,l,m	1	10597651	10598298	17_k,s
1	10652477	10652913	22_v	1	10852648	10853181	9,5,6_e,h,t,v,o,i,l,m
1	10852648	10867713	17_e,h,t,v,o,i,l,m	1	10867122	10867713	17_s
1	10929408	10931453	1,9,5,6_c,e,o,i,l,m	1	11153528	11156586	7,8,9_a,e,f,g,h,o,i,l,m
1	11153843	11156586	7,8,9_c,h	1	11174191	11174788	1,5,6_e,o,i,l,m
1	11298479	11299423	17_p	1	11351495	11352383	18,19_q
1	11407740	11410280	16,5,6_a,f,g,o,i,l,m	1	11521877	11522927	5,6_a,e,g,o,i,l,m
1	11521877	11526777	5,6_a,e,g,o,i,l,m	1	11525733	11526777	5,6_a,e,f,g,o,i,l,m
1	11617472	11618257	22_c,h,k,v,o,i,l,m	1	11647893	11648893	17_p
1	11647893	11649892	17_q	1	11649243	11649892	18,19_q
1	11669337	11670224	5,6_e,o,i,l,m	1	11699678	11701706	14_a,e,f,g,o,i,l,m
1	11699678	11708811	14_a,e,f,g,v,o,i,l,m	1	11706430	11708811	1,7,16,5,6_a,e,f,g,v,o,i,l,m
1	11720515	11732884	15_o,i,l,m	1	11732332	11732884	5,6_o,i,l,m
1	11782492	11783016	8,5,6_h,t,o,i,l,m	1	11803524	11804073	5,6_e,o,i,l,m
1	11803524	11806152	22_e,o,i,l,m	1	11805667	11806679	22_v,o,i,l,m
1	11805767	11806152	22_h,v,o,i,l,m	1	11980290	11982061	15,21_n
1	11980290	11988389	15,21_s	1	11980290	11992644	15,21_n
1	11990855	11992644	21_n	1	12084188	12097996	17_e,o,i,l,m

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
1	12096668	12097996	3,10,13_e,o,i,l,m	1	12097436	12097996	9,5,6_e,o,i,l,m
1	12136612	12137662	9,5,6_h,t,o,i,l,m	1	12183385	12183970	16,5,6_e,h,t,o,i,l,m
1	12240485	12240889	5,6_a,e,g,t,o,i,l,m	1	12327907	12328694	14_a,c,e,f,g,h,k,v,o,i,l,m
1	12327907	12330032	14_a,c,e,f,g,h,k,v,o,i,l,m	1	12364115	12365971	14_a,f,g,o,i,l,m
1	12457662	12458231	5,6,9_o,i,l,m	1	12562207	12564058	14_a,e,f,g,h,t,v,o,i,l,m
1	12618444	12619501	12_a,e,g,t,o,i,l,m	1	12618444	12621190	12_j
1	12618444	12636302	12_a,f,g	1	12618444	12636669	12_a,e,f,g,o,i,l,m
1	12618616	12619501	9,5,6_a,e,g,t,o,i,l,m	1	12618616	12621190	9,5,6_j
1	12618616	12636302	9,5,6_a,f,g	1	12618616	12636669	9,5,6_a,e,f,g,o,i,l,m
1	12853915	12855970	14_u,o,i,l,m	1	12873616	12874345	18,19_q
1	12873616	12875565	18,19_q	1	12873616	12876905	14_q
1	12875854	12876905	14_a,e,f,g,o,i,l,m	1	12992974	12993572	5,6_a,e,g,t,o,i,l,m
1	13015433	13017601	14_d	1	13015829	13017601	14_a,e,f,g,h,t,v,o,i,l,m
1	13058801	13059740	14_o,i,l,m	1	13058801	13060873	14_a,e,f,g,o,i,l,m
1	13058801	13064097	14_h	1	13064057	13065127	11_h,t,o,i,l,m
1	13064057	13065210	5,6_h,t,o,i,l,m	1	13160785	13162846	15_n
1	13212569	13215389	1,16_a,f,g,h,o,i,l,m	1	13262477	13263414	14_a,f,g,v,o,i,l,m
1	13314729	13315134	1,9,5,6_e,o,i,l,m	1	13358060	13367698	17_q
1	13358350	13367698	17_q	1	13366940	13367698	18,19_q
1	13372686	13387692	1,16_a,f,g,h	1	13372686	13388949	1,16_p
1	13426924	13439675	3,10,13_p	1	13474968	13477684	3,10,13_a,c,e,g,h,k,t
1	13474968	13481324	8,9,5,6_a,c,e,g,h,k,t	1	13474968	13482087	3,10,13_a,c,e,g,h,k,t
1	13477684	13481324	8,9,5,6_a,c,g,h,k,t	1	13477684	13482087	3,10,13_a,c,g,h,k,t
1	13477684	13497200	1,5,6_a,c,g,h,k,t	1	13477684	13497687	22_a,c,g,h,k,t
1	13480405	13481324	8,9,5,6_g	1	13480405	13482087	3,10,13_g
1	13480405	13497200	1,5,6_g	1	13480405	13497687	22_g
1	13496635	13497200	1,5,6_o,i,l,m	1	13496635	13497687	22_o,i,l,m
1	13549644	13567394	16,5,6_q	1	13566717	13567394	18,19_q
1	13566717	13584790	18,19_q	1	13583283	13584790	18,19_q
1	13602901	13604257	7,8,9_f,g	1	13625144	13625763	17_k,s
1	13809416	13809819	22_v,o,i,l,m	1	13809416	13810065	22_a,e,f,g
1	13812115	13812677	5,6_a,e,g,t,u,o,i,l,m	1	13871085	13871655	16,5,6_a,e,f,g,h,u,o,i,l,m
1	13880267	13880864	1,5,6_e,o,i,l,m	1	13880267	13883809	17_e,o,i,l,m
1	13882864	13883809	17_p,s	1	13980242	13983019	9,5,6_a,g
1	13980242	13990783	5,6_a,g	1	13981933	13983019	9,5,6_h,t,o,i,l,m
1	13981933	13990783	5,6_h,t,o,i,l,m	1	14015253	14015850	5,6_e,t,o,i,l,m
1	14015253	14016671	5,6_e,t,o,i,l,m	1	14298874	14300189	18,19_u,o,i,l,m
1	14298874	14302282	14_u,o,i,l,m	1	14298959	14300189	18,19_q
1	14298959	14302282	14_q	1	15644802	15645497	5,6_t
1	15941763	15942490	14_o,i,l,m	1	15941763	15943617	14_a,c,e,f,g,t,u,o,i,l,m
1	15941763	15947507	14_f	1	15962349	15963568	7,8,9_a,g
1	16071680	16079854	17_p,s	1	16072067	16079854	17_p,s
1	16076876	16079854	17_p,s	1	16140676	16147912	17_q
1	16144969	16147912	18,19_q	1	16161842	16163891	15_n
1	16262258	16264992	16,5,6_e,f,g,o,i,l,m	1	16319788	16329742	18,19_a,e,f,g,h,t,v,o,i,l,m
1	16327884	16329742	14_a,e,f,g,h,t,v,o,i,l,m	1	16329702	16345409	18,19_g
1	16396793	16398593	23_s,w	1	16396793	16403442	16,5,6_s,w
1	16402889	16403442	16,5,6_e,f,h,v,o,i,l,m	1	16424452	16426585	7,8_a,e,f,g
1	16480134	16480524	17_s	1	16542510	16549372	23_w
1	16545818	16549372	23_w	1	16593193	16595574	1,7,16,5,6_f,g,o,i,l,m
1	16701852	16704396	15_n	1	16758988	16761366	3,10,13_a,c,e,g,h,k,t
1	16794283	16794915	22_v	1	16794283	16806609	22_a,f,g,h,v,o,i,l,m
1	16794283	16810394	22_a,f,g,o,i,l,m	1	16808590	16810394	14_a,f,g,o,i,l,m
1	16808590	16822878	14_j	1	16808590	16825327	14_h,v,o,i,l,m
1	16809429	16810394	14_a,f,g,o,i,l,m	1	16809429	16822878	14_j
1	16809429	16825327	14_h,v,o,i,l,m	1	16809429	16828809	14_a,e,f,g,h,t,v,o,i,l,m
1	16824713	16825327	22_h,v,o,i,l,m	1	16824713	16828809	22_a,e,f,g,h,t,v,o,i,l,m
1	16824713	16830454	22_n	1	16824713	16831259	22_n
1	16826958	16828809	14_a,e,f,g,h,t,v,o,i,l,m	1	16826958	16830454	14_n
1	16826958	16831259	14_n	1	16828809	16830454	15_n
1	16828809	16831259	15_n	1	16861072	16861616	5,6_e,f,g,h,t,v,o,i,l,m
1	16861072	16862351	18,19_e,f,g,h,t,v,o,i,l,m	1	16861072	16866512	18,19_e,f,g,h,t,v,o,i,l,m
1	16861072	16867441	18,19_e,f,g,h,t,v,o,i,l,m	1	16861072	16879660	18,19_e,f,g,h,t,v,o,i,l,m
1	16862348	16866512	18,19_q	1	16862348	16867441	18,19_q
1	16862348	16879660	18,19_q	1	16863079	16866512	18,19_q
1	16863079	16867441	18,19_q	1	16863079	16879660	18,19_q
1	16906240	16907172	14_o,i,l,m	1	16906240	16908303	14_e,f,o,i,l,m
1	16968287	16970677	1,16,5,6_e,f,g,o,i,l,m	1	16968463	16970677	1,16,5,6_a,f,g,o,i,l,m
1	16999931	17001250	22_a,f,g,o,i,l,m	1	16999931	17001394	14_a,f,g,o,i,l,m
1	17000836	17001250	22_h,v,o,i,l,m	1	17000836	17001394	14_h,v,o,i,l,m
1	17104629	17113064	18,19_n	1	17104629	17113293	18,19_q
1	17104629	17114004	18,19_q	1	17104629	17116889	18,19_f
1	17104708	17113064	18,19_n	1	17104708	17113293	18,19_q

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
1	17104708	17114004	18,19_q	1	17104708	17116889	18,19_f
1	17109968	17113064	18,19_n	1	17109968	17113293	18,19_q
1	17109968	17114004	18,19_q	1	17109968	17116889	18,19_f
1	17110837	17113064	18,19_n	1	17110837	17113293	18,19_q
1	17110837	17114004	18,19_q	1	17110837	17116889	18,19_f
1	17111269	17113064	21_n	1	17111269	17113293	21_q
1	17111269	17114004	21_q	1	17111269	17116889	21_f
1	17116849	17117888	5,6_g,o,i,l,m	1	17334881	17335645	9,5,6_a,e,f,g,h,v,o,i,l,m
1	17407227	17408196	18,19_q	1	17407227	17408634	18,19_q
1	17504670	17505246	16,5,6_e,f,o,i,l,m	1	17553628	17555437	14_u
1	17607990	17608574	5,6_e,h,t,o,i,l,m	1	17903531	17904486	14_o,i,l,m
1	17903531	17905609	14_a,f,g,o,i,l,m	1	18024681	18026338	17_s
1	18290861	18291285	22_v	1	18303938	18304809	5,6_a,e,g,t,o,i,l,m
1	18324272	18325067	17_p	1	18635807	18636613	17_p
1	18635807	18637463	17_p	1	18636856	18637463	17_k,s
1	18769521	18770285	16,5,6_e,f,g,h,o,i,l,m	1	18793771	18794515	14_a,f,g,o,i,l,m
1	18793771	18795418	14_o,i,l,m	1	18793771	18796531	14_a,c,e,f,g,h,k,t,v,o,i,l,m
1	18794595	18795418	14_o,i,l,m	1	18794595	18796531	14_a,c,e,f,g,h,k,t,v,o,i,l,m
1	18796491	18808780	5,6_g	1	18984310	18985044	5,6_t
1	19058090	19059263	17_p	1	19058090	19060232	17_p
1	19058090	19066573	17_g,o,i,l,m	1	19058090	19067267	17_g,o,i,l,m
1	19058090	19068259	17_a,e,f,g,h,t,v,o,i,l,m	1	19065633	19066573	14_g,o,i,l,m
1	19065633	19067267	14_g,o,i,l,m	1	19065633	19068259	14_a,e,f,g,h,t,v,o,i,l,m
1	19088767	19090510	14_u	1	19149293	19152291	2_b
1	19175728	19177469	14_u	1	19238385	19241069	12_j
1	19330906	19332928	14_a,f,g,o,i,l,m	1	19330906	19341054	22_a,f,g,o,i,l,m
1	19331980	19332928	14_a,e,f,g,o,i,l,m	1	19331980	19341054	22_a,e,f,g,o,i,l,m
1	19395855	19397984	15_n	1	19406968	19407142	18,19_q
1	19468759	19470617	14_a,e,f,g,h,t,v,o,i,l,m	1	19669616	19670343	9,5,6_e,f,g,o,i,l,m
1	19728423	19740251	14_u	1	19745208	19745805	16,5,6_e,t,o,i,l,m
1	19766429	19767278	21_n	1	19766429	19773240	21_a,e,f,g,h,t,v,o,i,l,m
1	19771658	19773500	14_a,e,f,g,u,o,i,l,m	1	19772692	19773240	16,5,6_a,e,f,g,h,t,v,o,i,l,m
1	19945052	19949058	12_j	1	20170339	20171222	5,6_e,f,h,t,v,o,i,l,m
1	20170339	20176476	5,6_e,f,h,t,v,o,i,l,m	1	20175729	20176476	5,6_e,t,o,i,l,m
1	20209022	20209572	1,5,6_e,f,h,t,v,o,i,l,m	1	20512504	20513088	5,6_e,h,u,o,i,l,m
1	20517346	20518239	5,6,9_o,i,l,m	1	20707581	20708240	17_p
1	20823348	20824288	17_p	1	20975738	20977965	1,7,16,5,6_f,g,o,i,l,m
1	21107020	21108761	11_e,f,g,u,o,i,l,m	1	21107020	21108853	9,5,6_e,f,g,u,o,i,l,m
1	21131654	21132534	5,6_a,e,f,g,t,o,i,l,m	1	21263811	21264695	18,19_q
1	21339609	21340190	5,6_o,i,l,m	1	21391663	21392663	5,6_a,f,g,o,i,l,m
1	21431824	21432419	17_s	1	21618962	21619815	9,5,6_a,e,f,g,h,t,v,o,i,l,m
1	21742741	21743305	9,5,6_a,e,f,h,t,v,o,i,l,m	1	21791858	21792523	17_p
1	21828116	21828781	17_p	1	21875887	21881021	7,8,9_q
1	21922099	21922648	16,5,6_e,o,i,l,m	1	22195672	22196216	16,5,6_a,e,g,t,o,i,l,m
1	22453596	22455669	14_a,e,f,g,h,t,v,o,i,l,m	1	22624255	22625045	14_a,c,e,f,g,h,k,v,o,i,l,m
1	22624255	22625969	14_a,c,e,f,g,h,k,v,o,i,l,m	1	22625444	22625969	14_o,i,l,m
1	22832282	22832937	17_s	1	22855239	22857867	3,10,13_a,c,h,k,t,o,i,l,m
1	22856860	22857867	3,10,13_a,c,h,k,t,o,i,l,m	1	22995597	22996386	14_a,c,e,f,g,h,k,v,o,i,l,m
1	22995597	22997725	14_a,c,e,f,g,h,k,v,o,i,l,m	1	22996785	22997725	14_o,i,l,m
1	23128796	23129357	16,5,6_e,o,i,l,m	1	23177619	23178869	14_u
1	23217331	23219183	14_a,f,g,o,i,l,m	1	23557456	23558103	14_a,e,f,g,h,t,v,o,i,l,m
1	23879809	23880930	14_a,e,f,g,k,o,i,l,m	1	24536992	24538090	9,5,6_a,c,e,g,k,o,i,l,m
1	24722137	24722700	5,6_e,h,t,o,i,l,m	1	25361487	25362084	5,6_e,o,i,l,m
1	25966128	25966919	14_a,e,f,g,o,i,l,m	1	26173058	26175142	14_a,e,f,g,o,i,l,m
1	26310623	26311167	16,5,6_e,f,v,o,i,l,m	1	26514333	26514888	9,5,6_e,o,i,l,m
1	26884383	26884937	16,5,6_e,o,i,l,m	1	27244381	27244936	9,5,6_e,f,h,v,o,i,l,m
1	28315340	28316290	17_p	1	28754794	28755208	22_v
1	28896274	28896823	5,6_e,o,i,l,m	1	29083309	29083906	5,6_a,e,g,t,o,i,l,m
1	29249036	29249578	5,6_a,e,g,t,u,o,i,l,m	1	29779154	29780032	9,5,6_a,e,g,t,o,i,l,m
1	30255870	30256659	14_a,c,e,f,g,h,k,v,o,i,l,m	1	30255870	30257998	14_a,c,e,f,g,h,k,v,o,i,l,m
1	30255870	30260414	14_a,c,e,f,g,h,k,v,o,i,l,m	1	30257058	30257998	14_f,o,i,l,m
1	30257058	30260414	14_f,o,i,l,m	1	30258034	30260414	14_a,e,f,g,o,i,l,m
2	126927	128151	7,16_g	2	238046	238123	15,21_n
2	238046	241922	15,21_q	2	240845	241922	18,19_q
2	499984	502058	14_a,f,g,o,i,l,m	2	499984	511337	14_v
2	510987	511337	22_v	2	542307	544163	14_a,e,f,g,k,u,o,i,l,m
2	707441	723165	14_h	2	715662	716838	17_p
2	715662	728072	23_p	2	721903	723165	17_h
2	723125	728072	23_h,t,o,i,l,m	2	725859	728072	23_w
2	791410	793900	7,8,9_v	2	794446	795635	17_p
2	825443	826272	5,6_e,o,i,l,m	2	979824	980613	14_a,e,f,g,h,o,i,l,m
2	1038046	1038909	5,6_a,e,f,g,h,t,u,v,o,i,l,m	2	1038046	1049404	17_a,e,f,g,h,t,u,v,o,i,l,m
2	1048325	1049404	17_p	2	1075836	1076403	17_s
2	1087659	1088933	18,19_q	2	1087659	1089963	18,19_q

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
2	1087659	1092283	18,19_q	2	1088986	1089963	18,19_q
2	1088986	1092283	18,19_q	2	1089388	1089963	18,19_q
2	1089388	1092283	18,19_q	2	1089961	1092283	18,19_q
2	1090121	1092283	18,19_q	2	1090279	1092283	18,19_q
2	1126464	1127487	17_p	2	1212983	1219988	15_r
2	1233681	1234504	3,10,13_h	2	1280864	1281518	17_s
2	1286166	1286933	22_v	2	1302354	1303109	1,9,5,6_e,f,g,o,i,l,m
2	1321967	1322852	9,5,6_e,t,o,i,l,m	2	1347209	1348072	9,5,6_a,e,g,t,o,i,l,m
2	1367472	1368498	9,5,6_a,c,e,f,g,k,t,o,i,l,m	2	1367472	1374277	11_a,c,e,f,g,k,t,o,i,l,m
2	1367472	1384410	5,6_a,c,e,f,g,k,t,o,i,l,m	2	1383550	1384410	5,6_a,e,g,t,o,i,l,m
2	1409584	1410151	5,6_o,i,l,m	2	1441891	1446897	17_f,g,h,o,i,l,m
2	1458844	1460921	14_u,o,i,l,m	2	1552028	1552858	5,6_h,t,o,i,l,m
2	1596916	1597475	5,6_e,t,o,i,l,m	2	1669788	1670330	5,6,16_o,i,l,m
2	1691161	1692148	17_p	2	1788917	1789902	5,6_h,t,o,i,l,m
2	1804014	1821389	4_d	2	1847205	1858451	7,8_w
2	1847641	1858451	7,8,9_w	2	1902860	1903641	18,19_q
2	1939906	1948488	17_a,f,g	2	1940154	1948488	17_c,g,h,t
2	1947591	1948488	17_p	2	1947591	1963668	17_p
2	1962765	1963668	17_p,s	2	1987731	1988685	17_p
2	2016981	2020884	18,19_n	2	2018362	2020884	15_n
2	2058072	2059184	17_p,s	2	2117334	2121242	4_a,e,f,g,k,t,o,i,l,m
2	2117334	2129288	8,5,6_a,e,f,g,k,t,o,i,l,m	2	2117334	2130303	18,19_a,e,f,g,k,t,o,i,l,m
2	2118110	2121242	4_d	2	2118110	2129288	8,5,6_d
2	2118110	2130303	18,19_d	2	2138817	2139381	1,5,6_e,f,h,t,v,o,i,l,m
2	2174239	2175666	21_n	2	2219190	2219855	9,5,6_a,e,g,o,i,l,m
2	2240577	2241227	17_c,k,s,w	2	2240577	2256585	18,19_c,k,s,w
2	2246165	2256585	18,19_b	2	2246165	2262401	1,9,5,6_b
2	2261415	2262401	1,9,5,6_e,f,g,t,o,i,l,m	2	2297109	2297757	17_k,s
2	2297109	2299661	17_e,o,i,l,m	2	2299064	2299661	5,6_e,o,i,l,m
2	2320059	2320930	5,6_a,e,f,g,h,t,v,o,i,l,m	2	2320059	2325721	5,6_f
2	2320059	2326983	5,6_f,g,o,i,l,m	2	2320059	2327232	5,6_f,g
2	2338945	2340269	18,19_q	2	2442262	2443214	5,6_e,f,g,o,i,l,m
2	2517549	2520193	1,7,16_c	2	2629305	2630941	7,8_a,e,f,g
2	2653112	2659258	7,8_g	2	2734775	2735211	22_v
2	2738275	2739754	23_s,w	2	2820417	2820852	22_v
2	2843834	2850690	17_e,o,i,l,m	2	2843834	2852357	17_s
2	2844639	2850690	8,5,6_e,o,i,l,m	2	2844639	2852357	8,5,6_s
2	2911725	2913287	7,8_a,e,f,g,o,i,l,m	2	2911971	2913287	7,8_a,f,g
2	2961259	2962331	17_g	2	2962291	2962801	14_a,e,f,g,t,o,i,l,m
2	3191174	3191784	17_s	2	3191174	3198498	9,5,6_s
2	3192306	3193960	16,5,6_n	2	3192306	3195996	16,5,6_n
2	3192306	3212271	16,5,6_a,f,g,o,i,l,m	2	3192306	3212362	16,5,6_g
2	3192429	3193960	15_n	2	3192429	3195996	15_n
2	3192429	3212271	15_a,f,g,o,i,l,m	2	3192429	3212362	15_g
2	3197360	3198498	9,5,6_h,v,o,i,l,m	2	3197360	3211631	22_h,v,o,i,l,m
2	3197360	3213350	14_h,v,o,i,l,m	2	3197615	3198498	9,5,6_e,o,i,l,m
2	3197615	3211631	22_e,o,i,l,m	2	3197615	3213350	14_e,o,i,l,m
2	3212231	3213350	14_f,g	2	3212231	3218098	22_f,g
2	3212322	3213350	14_a,e,f,g,h,t,v,o,i,l,m	2	3212322	3218098	22_a,e,f,g,h,t,v,o,i,l,m
2	3598084	3598726	17_s	2	3727343	3728601	7,8,9,16_a,e,f,g,o,i,l,m
2	3937053	3940463	15_n	2	4031284	4032072	14_a,c,e,f,g,h,k,v,o,i,l,m
2	4092156	4093662	18,19_q	2	4092156	4094540	18,19_q
2	4093662	4094540	18,19_q	2	4153829	4155183	8,9,5,6_a,f,g,o,i,l,m
2	4314360	4314784	22_v	2	4357213	4358656	18,19_g
2	4358616	4360640	9,5,6_a,e,f,g,v,o,i,l,m	2	4358616	4362792	17_a,e,f,g,v,o,i,l,m
2	4358616	4368261	18,19_a,e,f,g,v,o,i,l,m	2	4361881	4362792	17_p
2	4361881	4368261	18,19_p	2	4457892	4458619	17_p
2	4478897	4479546	17_s	2	4478897	4483362	17_n
2	4481504	4483362	15_n	2	4658105	4661376	7,8_a,e,g,h,k,o,i,l,m
2	4658773	4661289	1,16,5,6_a,c,h,t	2	4753826	4755843	18,19_q
2	4754021	4755843	18,19_q	2	4765183	4765777	5,6_e,o,i,l,m
2	4765183	4766954	17_e,o,i,l,m	2	4766145	4766954	17_p
2	4791907	4801035	2_u	2	4803429	4804420	9,5,6_e,f,o,i,l,m
2	4948582	4952253	22_c	2	5017169	5017733	1,5,6_a,e,g,k,t,o,i,l,m
2	5044833	5045394	1,9,5,6_a,e,f,g,h,t,v,o,i,l,m	2	5055631	5056760	18,19_q
2	5080143	5087968	3,10,13_f	2	5354208	5355206	9,5,6_h,t
2	5431529	5433456	23_s,w	2	5472005	5474085	14_u
2	5614526	5615673	18,19_q	2	5654748	5655421	18,19_q
2	5843829	5844646	17_p,s	2	6010847	6011886	5,6_a,c,e,t,o,i,l,m
2	6010847	6013302	3,10,13_a,c,e,t,o,i,l,m	2	6010847	6020525	20_a,c,e,t,o,i,l,m
2	6066670	6077867	2_b	2	6090825	6094816	15_c,g
2	6090825	6108457	14_c,g	2	6091172	6092407	14_u
2	6091172	6109808	14_u	2	6092970	6094816	15_n
2	6092970	6108457	14_n	2	6093724	6094816	15_n

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
2	6093724	6108457	14_n	2	6106612	6109808	14_u
2	6107314	6108457	14_a,e,f,g,h,k,u,o,i,l,m	2	6206514	6215590	12_h,k
2	6212701	6215590	3,10,13_h,k	2	6213677	6214232	16,5,6_a,e,f,g,h,t,v,o,i,l,m
2	6269029	6270285	17_p	2	6269029	6279178	17_a,f,g
2	6304288	6305069	18,19_q	2	6325950	6328356	23_w
2	6334748	6352095	15_n	2	6378297	6381304	3,10,13_c,g,h,k
2	6418231	6419233	1,9,5,6_e,f,g,t,o,i,l,m	2	6476177	6476873	15_n
2	6528510	6530701	3,10,13_h	2	6636353	6636954	17_s
2	6664571	6665545	5,6_g,t,o,i,l,m	2	6664571	6671202	5,6_a,c,e,f,g,h,k,v,o,i,l,m
2	6668043	6671202	14_a,c,e,f,g,h,k,v,o,i,l,m	2	6681197	6690194	17_p
2	6685533	6687910	1,7,16,5,6_f,g,o,i,l,m	2	6689255	6690194	17_p
2	6859748	6862117	1,7,16,5,6_a,e,f,g,v,o,i,l,m	2	6908273	6909003	9,5,6_a,e,g,t,o,i,l,m
2	6930373	6932458	14_a,e,f,g,o,i,l,m	2	6953237	6953673	22_v
2	7029378	7030871	7,8,9_a,e,f,g,o,i,l,m	2	7030831	7031367	18,19_f
2	7100902	7101634	5,6_e,o,i,l,m	2	7100902	7120009	5,6_a,c,e,f,g,h,k,v,o,i,l,m
2	7118866	7120009	14_a,c,e,f,g,h,k,v,o,i,l,m	2	7160877	7161817	14_o,i,l,m
2	7160877	7176605	14_o,i,l,m	2	7175312	7176605	14_f,o,i,l,m
2	7175312	7184674	5,6_f,o,i,l,m	2	7175312	7185505	5,6_f,o,i,l,m
2	7183822	7184674	5,6_e,t,o,i,l,m	2	7183822	7185505	5,6_e,t,o,i,l,m
2	7184672	7185505	5,6_a,e,f,g,t,v,o,i,l,m	2	7185501	7186096	5,6_a,e,g,t,o,i,l,m
2	7347463	7348023	16,5,6_e,f,h,v,o,i,l,m	2	7347463	7358628	16,5,6_a,e,f,g,h,t,v,o,i,l,m
2	7354868	7358628	17_a,e,f,g,h,t,v,o,i,l,m	2	7357736	7358628	5,6_a,e,f,g,h,t,v,o,i,l,m
2	7446035	7446804	1,9,5,6_h,t	2	7573631	7574706	9,5,6_a,c,e,f,g,k,t,o,i,l,m
2	7657442	7658062	22_v	2	7736377	7737651	22_q
2	7820484	7821772	14_a,f,g,o,i,l,m	2	8035448	8037831	5,6_a,e,f,g,h,k,o,i,l,m
2	8035448	8053350	1,9,5,6_a,e,f,g,h,k,o,i,l,m	2	8052774	8053350	1,9,5,6_e,k,o,i,l,m
2	8052774	8069181	17_e,k,o,i,l,m	2	8179917	8180781	9,5,6_e,t,o,i,l,m
2	8243344	8244498	17_p,s	2	8861453	8862342	18,19_q
2	8900618	8902246	20_r	2	8914532	8915076	5,6_e,f,g,h,t,v,o,i,l,m
2	9036888	9037755	9,5,6_a,e,g,t,o,i,l,m	2	9098219	9099202	9,5,6_t,o,i,l,m
2	9232556	9233104	9_e,o,i,l,m	2	9232556	9236258	16,5,6_e,o,i,l,m
2	9235717	9236258	16,5,6_a,e,g,t,o,i,l,m	2	9440171	9440648	16,5,6_a,e,g,t,o,i,l,m
2	9591687	9592434	9,5,6_e,f,h,t,v,o,i,l,m	2	9619101	9620020	5,6_a,e,f,g,o,i,l,m
2	9843571	9844098	8,5,6_h,t,o,i,l,m	2	9867035	9867593	16,5,6_e,f,h,o,i,l,m
2	10025153	10027535	1,16,5,6_a,f,g,o,i,l,m	2	10043614	10047099	12_j
2	10077180	10078043	5,6_e,o,i,l,m	2	10262066	10263126	17_s
2	10262066	10264272	3,10,13_s	2	10358652	10359540	18,19_q
2	10413581	10414173	22_h,v,o,i,l,m	2	10472308	10473535	16_c
2	10474356	10476158	14_u	2	11003261	11004120	5,6_a,c,e,f,g,h,k,t,v,o,i,l,m
2	11135890	11136295	22_k,v	2	11292193	11294238	14_a,e,f,g,h,t,v,o,i,l,m
2	11293274	11294238	14_o,i,l,m	2	11404563	11406454	14_u
2	11562480	11563379	18,19_q	2	11604370	11604956	16_e,o,i,l,m
2	11915953	11916796	5,6_e,o,i,l,m	2	12100559	12101501	14_o,i,l,m
2	12150781	12151854	17_p	2	12561112	12563279	14_u
2	12563731	12564280	5,6_e,o,i,l,m	2	12808071	12809030	14_o,i,l,m
2	12808071	12810170	14_a,e,f,g,o,i,l,m	2	13071584	13075465	12_j
2	13108887	13109436	16,5,6_a,e,f,g,h,t,v,o,i,l,m	2	13273992	13274947	14_o,i,l,m
2	13273992	13276081	14_e,f,o,i,l,m	2	13342901	13345153	14_a,e,f,g,o,i,l,m
2	13893008	13893555	5,6_f,o,i,l,m	2	13913744	13915797	14_a,e,f,g,o,i,l,m
2	14578973	14579824	5,6_e,f,o,i,l,m	2	15141304	15141850	5,6_e,f,h,t,v,o,i,l,m
2	15378177	15378958	18,19_q	2	17260236	17262310	14_f,o,i,l,m
2	17261365	17262310	14_o,i,l,m	2	17306500	17307232	18,19_q
2	17306500	17307512	18,19_q	2	17999621	18000164	5,6_e,f,v,o,i,l,m
2	17999621	18000227	16,5,6_e,f,v,o,i,l,m	2	19202480	19203826	14_u
2	19275667	19276267	22_v	2	19420986	19421723	5,6_h,t
2	19535011	19535918	18,19_q	2	19630172	19630715	5,6_e,o,i,l,m
3	538788	539202	22_v	3	1283143	1283997	17_p
3	1341718	1342247	5,6_c,h,t,o,i,l,m	3	1571386	1572131	14_a,c,e,f,g,h,k,v,o,i,l,m
3	2628489	2630252	14_u	3	2647198	2648519	18,19_q
3	2919936	2920519	5,6_h,t,o,i,l,m	3	3080030	3080595	16,5,6_e,o,i,l,m
3	3716348	3717217	14_a,e,f,g,v,o,i,l,m	3	3987686	3988443	18,19_q
3	4274937	4289116	1,7,16,5,6_a,e,f,g,h,o,i,l,m	3	4551556	4552322	16,5,6_a,e,f,g,h,v,o,i,l,m
3	4579244	4579712	9,5,6_f,o,i,l,m	3	4579244	4585303	9,5,6_a,e,g,h,t,o,i,l,m
3	4579672	4585261	17_g	3	4584763	4585303	5,6_a,e,g,h,t,o,i,l,m
3	4795963	4796835	18,19_q	3	4795963	4797021	18,19_q
3	4977275	4978019	5,6_e,t,o,i,l,m	3	5379344	5380237	18,19_q
3	5406658	5407231	1,7,8_f,g	3	5590765	5591642	9,5,6_e,o,i,l,m
3	5622037	5624129	14_u	3	5865870	5866434	5,6_a,e,g,t,o,i,l,m
3	6052701	6054920	15_n	3	6728931	6729504	5,6_a,e,f,g,h,t,v,o,i,l,m
3	6940605	6941489	9,5,6_e,f,v,o,i,l,m	3	6991501	6992369	18,19_q
3	6991501	6992791	18,19_q	3	6991501	6993214	18,19_q
3	7291171	7291973	18,19_q	3	7318508	7319275	1,9,5,6_a,e,f,g,h,v,o,i,l,m
3	7658942	7659527	16,5,6_e,h,o,i,l,m	3	7765528	7766849	17_p
3	7766069	7766623	16,5,6_e,o,i,l,m	3	7936991	7937577	16,5,6_e,h,o,i,l,m

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
3	7980090	7980639	16,5,6_a,e,g,t,o,i,l,m	3	7980090	7996642	14_a,e,g,t,o,i,l,m
3	7981215	7982106	5,6_a,e,f,g,o,i,l,m	3	7981215	7994876	5,6_g
3	7994836	7996642	14_a,e,f,g,o,i,l,m	3	8596111	8596973	5,6_e,t,o,i,l,m
3	8596111	8614552	7,8_e,t,o,i,l,m	3	8613588	8614552	7,8_a,f,g
3	8644749	8645345	5,6_a,e,f,g,h,v,o,i,l,m	3	8772154	8772892	1,9,5,6_h,t,o,i,l,m
3	8874904	8875681	17_p	3	8874904	8884071	17_p
3	8883285	8884071	17_p	3	9257757	9258725	17_p
3	9541696	9542484	5,6_a,e,f,g,h,v,o,i,l,m	3	9559440	9560919	14_u,o,i,l,m
3	9767594	9768865	14_a,f,g,o,i,l,m	3	9768920	9769339	22_h
3	9891286	9892211	14_a,e,f,g,v,o,i,l,m	3	10050151	10050786	17_s
3	10083297	10084064	16,5,6_a,e,f,g,h,v,o,i,l,m	3	10177909	10179013	17_p
3	10202900	10203856	17_p	3	10321649	10322677	14_a,e,f,g,h,t,v,o,i,l,m
3	10406633	10407195	9,5,6_e,o,i,l,m	3	10406633	10418908	9,5,6_p
3	10418103	10418908	17_p	3	10510757	10512990	9,5,6_a,e,f,g,v,o,i,l,m
3	10557369	10558582	14_a,f,g,o,i,l,m	3	10575588	10587211	15_a,e,f,g,o,i,l,m
3	10643525	10645021	18,19_q	3	10682629	10683188	16,5,6_a,e,f,g,h,t,v,o,i,l,m
3	10709862	10710415	16,5,6_e,f,h,v,o,i,l,m	3	10719031	10719581	5,6_e,o,i,l,m
3	10752353	10753601	20_r	3	10752353	10759816	9,5,6_r
3	10759239	10759816	9,5,6_e,o,i,l,m	3	10792350	10792785	22_v
3	10821723	10822682	14_f,o,i,l,m	3	10826380	10826998	17_s
3	10826380	10839299	17_n	3	10826380	10844594	17_q
3	10837539	10839299	21_n	3	10837539	10844594	21_q
3	10842787	10844594	18,19_q	3	10863638	10867285	14_u
3	10901999	10902882	9,5,6_e,o,i,l,m	3	10901999	10907557	9,5,6_e,g,t,o,i,l,m
3	10906555	10907557	5,6_e,g,t,o,i,l,m	3	10909241	10909710	22_v
3	10978828	10981421	2_b	3	10998419	10999040	17_p
3	10998419	11002445	23_p	3	10999839	11002445	23_q
3	11000195	11002445	23_w	3	11054245	11073493	2_n
3	11091677	11092239	5,6_e,t,o,i,l,m	3	11111985	11113301	7,8,9_a,e,f,g,h,o,i,l,m
3	11111985	11113551	7,8,9_a,e,f,g,o,i,l,m	3	11133439	11133981	16,5,6_e,f,h,t,v,o,i,l,m
3	11168038	11168629	17_s	3	11168038	11177375	9,5,6_s
3	11171061	11177375	9,5,6_e,f,o,i,l,m	3	11171061	11189703	9_e,f,o,i,l,m
3	11189113	11189703	9_a,c,e,g,h,k,s,t,o,i,l,m	3	11189113	11204009	15_a,c,e,g,h,k,s,t,o,i,l,m
3	11201049	11205256	17_p	3	11201691	11204009	15_n
3	11201861	11205256	14_p	3	11244730	11247825	15_n
3	11245403	11247825	15_n	3	11287700	11288711	9,5,6_e,o,i,l,m
3	11306601	11309560	7,8_g	3	11306601	11309692	7,8_f,g
3	11306601	11316787	7,8_e,f,h,o,i,l,m	3	11316244	11316787	1,9,5,6_e,f,h,o,i,l,m
3	11346999	11348885	15_n	3	11377558	11384467	2_b
3	11458038	11458782	9,5,6_e,t,o,i,l,m	3	11537751	11542381	9,5,6_h,t,o,i,l,m
3	11537751	11554419	17_h,t,o,i,l,m	3	11538323	11541947	14_n
3	11540368	11541947	15_n	3	11553442	11554419	17_p
3	11696670	11698356	7,8,9_a,e,f,g	3	11765029	11766325	7,8,9,16_a,f,g,h
3	11765029	11768348	7,8,9,16_a,f,g,o,i,l,m	3	11826287	11828507	15_n
3	11826287	11828792	15,21_n	3	11865709	11866782	17_p
3	11930411	11932676	7,8,9_a,f,g,o,i,l,m	3	11961880	11962772	17_p
3	11961880	11969753	2_p	3	11965345	11969753	2_b
3	11970132	11971095	14_a,e,f,g,u,o,i,l,m	3	12149225	12160591	2_b
3	12662096	12664477	16_a,e,f,g,v,o,i,l,m	3	12801093	12817205	9,5,6_a,e,f,g,h
3	12885503	12886556	18,19_q	3	12901612	12907241	2_b
3	13261564	13266234	2_b	3	13261564	13274289	2_q
3	13273242	13274289	18,19_q	3	13299026	13300073	18,19_q
3	13300855	13307523	2_b	3	13300855	13311680	5,6_b
3	13300855	13317816	2_b	3	13310508	13311680	5,6_b
3	13310508	13317816	2_b	3	13331147	13332294	14_a,c,e,f,g,h,k,t,v,o,i,l,m
3	13487652	13493193	4_d	3	13834327	13834934	17_s
3	13834327	13852087	17_h	3	13852047	13854357	3,10,13_a,c,g,h,k,t,o,i,l,m
3	13852233	13854357	3,10,13_c	3	14068110	14068672	5,6_f,o,i,l,m
3	14153985	14169036	17_k,s	3	14283645	14297958	18,19_q
3	14283943	14296438	16,5,6_a,e,g,k,o,i,l,m	3	14308028	14310841	17_p
3	14530898	14546231	3,10,13_c,h,k	3	14570067	14570843	17_p
3	14796181	14796536	22_v	3	14847911	14848413	18,19_q
3	14914225	14916174	1,7,16,5,6_c,h,t	3	15031814	15032873	9,5,6_a,e,g,o,i,l,m
3	15099715	15102497	2_b	3	15199358	15199968	17_s
3	15289600	15304550	14_u,o,i,l,m	3	15349833	15360713	18,19_a,c,h,k,t
3	15495517	15496737	14_a,e,f,g,o,i,l,m	3	15514676	15516469	21_n
3	15577120	15577733	18,19_q	3	15694252	15696636	2_b
3	15746920	15760851	9,5,6_p	3	15746920	15761061	21_p
3	15751307	15752370	17_p	3	15759965	15760851	9,5,6_f
3	15759965	15761061	21_f	3	15759965	15774609	18,19_f
3	15794142	15794578	22_v,o,i,l,m	3	15823949	15825477	5,6_e,t,o,i,l,m
3	15854968	15857089	15_n	3	15884947	15885722	17_p
3	15884947	15895284	9,5,6_p	3	15929080	15929677	16,5,6_a,e,g,t,o,i,l,m
3	15937432	15938468	17_p	3	15937432	15949212	17_o,i,l,m

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
3	15937432	15950357	17_v,o,i,l,m	3	15946451	15949212	3,10,13_o,i,l,m
3	15946451	15950357	3,10,13_v,o,i,l,m	3	15948273	15949212	14_o,i,l,m
3	15948273	15950357	14_v,o,i,l,m	3	15987765	15989120	18,19_q
3	16072502	16073781	11_e,f,g,o,i,l,m	3	16072502	16073873	9,5,6_e,f,g,o,i,l,m
3	16096418	16100335	15_q	3	16099202	16100335	18,19_q
3	16099202	16118538	18,19_n	3	16110491	16111754	11_e,f,g,o,i,l,m
3	16110491	16111846	5,6_e,f,g,o,i,l,m	3	16183329	16197165	17_o,i,l,m
3	16196359	16197165	17_p	3	16233611	16234744	18,19_q
3	16233611	16246695	16,5,6_q	3	16235221	16236567	9,5,6_e,f,g,o,i,l,m
3	16235313	16236567	11_e,f,g,o,i,l,m	3	16245647	16246695	16,5,6_e,f,h,t,u,v,o,i,l,m
3	16277599	16278207	17_s	3	16277599	16287861	5,6_s
3	16278398	16290332	21_p,s	3	16286819	16287861	5,6_e,o,i,l,m
3	16288361	16290332	17_p,s	3	16416181	16416736	9,5,6_e,f,h,t,v,o,i,l,m
3	16416181	16419177	9,5,6_a,e,f,g,o,i,l,m	3	16416181	16422041	9,5,6_n
3	16416181	16431438	9,5,6_f,o,i,l,m	3	16420833	16422041	15,21_n
3	16420833	16431438	15,21_f,o,i,l,m	3	16430129	16431438	14_f,o,i,l,m
3	16461771	16463564	14_a,e,f,g,o,i,l,m	3	16496293	16496853	1,9,5,6_a,c,e,g,t,o,i,l,m
3	16530733	16533229	2_b	3	16564341	16565432	9,5,6_h,t,o,i,l,m
3	16564341	16573461	5,6,9_o,i,l,m	3	16564341	16574571	9,5,6_a,c,e,f,g,h,k,t,v,o,i,l,m
3	16572639	16573461	14_o,i,l,m	3	16572639	16574571	14_a,c,e,f,g,h,k,t,v,o,i,l,m
3	16618065	16620330	15,21_n	3	16680721	16689487	20_e,o,i,l,m
3	16688601	16689487	9,5,6_e,o,i,l,m	3	16738552	16739646	9,5,6_h,t
3	16862390	16862826	22_v,o,i,l,m	3	16862390	16865229	15_v,o,i,l,m
3	16863224	16865229	15_n	3	16913317	16914185	9,5,6_e,o,i,l,m
3	17003550	17006852	12_j	3	17044239	17045673	18,19_q
3	17072371	17074213	14_a,e,f,g,h,t,v,o,i,l,m	3	17072371	17081806	16,5,6_a,e,f,g,h,t,v,o,i,l,m
3	17079721	17081806	16,5,6_a,e,f,g,v,o,i,l,m	3	17112109	17112741	17_s
3	17112109	17128605	17_s	3	17127698	17128605	17_p
3	17153432	17155474	14_a,e,f,g,h,t,v,o,i,l,m	3	17153432	17161744	17_a,e,f,g,h,t,v,o,i,l,m
3	17220557	17221536	5,6_o,i,l,m	3	17286728	17288798	14_a,e,f,g,h,t,v,o,i,l,m
3	17310135	17310924	14_a,c,e,f,g,h,k,v,o,i,l,m	3	17310135	17312262	14_a,c,e,f,g,h,k,v,o,i,l,m
3	17311323	17312262	14_o,i,l,m	3	17406964	17408284	14_u
3	17406964	17413536	16,5,6_u	3	17408525	17411018	21_n
3	17412986	17413536	16,5,6_a,e,g,t,o,i,l,m	3	17480771	17482112	20_r
3	17740925	17742216	14_a,t,o,i,l,m	3	18029766	18030911	14_a,c,e,f,g,h,k,v,o,i,l,m
3	18157507	18158092	5,6_o,i,l,m	3	18183259	18183776	5,6_e,o,i,l,m
3	18423139	18424283	14_a,e,f,g,o,i,l,m	3	18506132	18507016	5,6_e,o,i,l,m
3	18595721	18596085	14_f,o,i,l,m	3	18644925	18645485	5,6_e,o,i,l,m
3	18854470	18855684	20_r	3	18896605	18898497	1,5,6_o,i,l,m
3	18959343	18960344	5,6_e,k,t,o,i,l,m	3	18987700	18988135	22_v,o,i,l,m
3	19447849	19448411	5,6_e,o,i,l,m	3	19531619	19532204	5,6_e,h,t,o,i,l,m
3	19531619	19540741	5,6_e,t,o,i,l,m	3	19540001	19540741	9,5,6_e,t,o,i,l,m
3	19547466	19548710	11_a,e,f,g,t,o,i,l,m	3	19547466	19548793	9,5,6_a,e,f,g,t,o,i,l,m
3	19659121	19659557	22_v	3	19750735	19751317	16,5,6_a,e,f,g,h,t,v,o,i,l,m
3	20494023	20496116	14_u	3	20494023	20498118	14_u
3	20780578	20781442	5,6_a,e,f,g,h,t,v,o,i,l,m	3	20797378	20799425	14_a,f,g,o,i,l,m
3	20886576	20887459	18,19_q	3	21483100	21483708	22_v
3	21563078	21563634	5,6_a,e,h,t,v,o,i,l,m	3	21705147	21705887	1,7,8,9,5,6_t
3	22129590	22131633	14_a,e,f,g,o,i,l,m	3	22226301	22226933	17_s
3	22579525	22581707	22_v	4	312202	314268	14_u
4	370788	372085	14_a,e,f,g,h,o,i,l,m	4	563509	564260	15_n
4	563509	572758	9,5,6_n	4	572197	572758	9,5,6_e,o,i,l,m
4	572197	591672	18,19_e,o,i,l,m	4	683213	684357	14_a,c,e,f,g,h,o,i,l,m
4	792040	793141	5,6_e,f,o,i,l,m	4	963674	964406	5,6_a,e,f,g,o,i,l,m
4	963674	965798	14_a,e,f,g,o,i,l,m	4	963863	964406	5,6_e,o,i,l,m
4	963863	965798	14_e,o,i,l,m	4	999920	1002006	14_a,e,f,g,o,i,l,m
4	1001051	1002006	14_o,i,l,m	4	1095275	1096866	18,19_q
4	1100279	1101040	5,6,9_o,i,l,m	4	1198713	1199584	9,5,6_a,e,g,t,o,i,l,m
4	1280625	1283009	16,5,6_a,f,g,o,i,l,m	4	1282969	1302899	18,19_f,g
4	1349585	1351655	14_a,f,g,o,i,l,m	4	1480680	1481327	7,8,9,5,6_a,f,g
4	1521255	1521816	16,5,6_a,e,g,t,o,i,l,m	4	1748835	1758876	3,10,13_c,f,g,h,k
4	1748835	1759023	3,10,13_c,f,g,h,k	4	1777274	1778169	18,19_q
4	1884301	1887148	1,7,8_a,f,g,h	4	1956609	1957700	9,5,6_a,c,e,f,g,k,t,o,i,l,m
4	2084369	2088366	16,5,6_f,o,i,l,m	4	2088655	2089814	18,19_q
4	2120177	2121349	5,6_o,i,l,m	4	2316087	2316521	22_v,o,i,l,m
4	2467970	2470263	23_w	4	2520882	2528029	14_q
4	2530681	2550356	8,9,5,6_q	4	2553546	2553946	9,5,6_e,o,i,l,m
4	2648906	2650845	23_w	4	2987143	2987909	5,6_a,e,f,g,h,v,o,i,l,m
4	3294252	3300993	7,8,9_g	4	3801486	3811728	23_w
4	3894016	3896909	1_a,e,f,g,h,k,o,i,l,m	4	3894016	3897754	1_e,h,u,o,i,l,m
4	3896869	3898126	14_f	4	3897169	3897754	1,9,5,6_e,h,u,o,i,l,m
4	4077215	4081131	5,6_e,o,i,l,m	4	4086836	4088667	2_b
4	4097211	4109850	23_b	4	4146123	4147061	14_o,i,l,m

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
4	4146123	4148178	14_a,e,f,g,h,t,v,o,i,l,m	4	4221540	4232536	2_b
4	4606225	4623085	5,6,16_o,i,l,m	4	4621818	4623085	5,6,9,8_o,i,l,m
4	4635138	4650543	18,19_v,o,i,l,m	4	4650107	4650543	22_v,o,i,l,m
4	4661572	4662689	3,10,13_a,c,h,k,t,o,i,l,m	4	4661572	4663902	3,10,13_a,c,h,k,t,o,i,l,m
4	4685635	4686071	22_v,o,i,l,m	4	4800959	4801843	18,19_q
4	4875410	4876057	22_q	4	4875410	4881072	18,19_q
4	4904289	4907102	3,10,13_h	4	4904289	4910116	3,10,13_h
4	4905414	4906326	22_v	4	4905414	4909344	22_v
4	4948252	4949062	17_p	4	4948252	4955160	18,19_p
4	5110646	5111687	5,6_o,i,l,m	4	5193724	5194591	18,19_q
4	5227436	5229704	15_n	4	5227953	5229704	21_n
4	5273983	5281178	1,7,8,9,16_p,s	4	5276482	5281178	3,10,13_p,s
4	5280201	5281178	17_p,s	4	5353745	5355719	14_u
4	5354111	5355719	14_u	4	5354473	5355719	14_u
4	5391717	5392279	1,5,6_e,o,i,l,m	4	5433090	5441436	5,6_f,g
4	5433222	5441436	16,5,6_f,g	4	5441396	5443567	16,5,6_a,f,g,o,i,l,m
4	5441396	5451092	18,19_a,f,g,o,i,l,m	4	5516921	5518896	17_p
4	5539314	5541385	14_a,f,g,o,i,l,m	4	5539314	5546451	15_a,f,g,o,i,l,m
4	5539314	5548708	15_a,f,g,o,i,l,m	4	5540445	5541385	14_f,o,i,l,m
4	5540445	5546451	15_f,o,i,l,m	4	5540445	5548708	15_f,o,i,l,m
4	5541418	5546451	15_n	4	5541418	5548708	15_n
4	5546716	5548708	15_n	4	5604457	5605412	17_p
4	5851690	5854645	21_n	4	5932510	5937002	3,10,13_h
4	6033542	6034615	17_p	4	6033542	6045992	17_p
4	6045088	6045992	17_p	4	6115338	6115899	16,5,6_e,o,i,l,m
4	6140443	6140942	17_p,s	4	6140443	6159503	17_p,s
4	6157517	6159503	17_p,s	4	6157517	6169329	17_h,t,o,i,l,m
4	6157875	6166460	14_e,f,g,o,i,l,m	4	6157875	6169225	5,6_e,f,g,o,i,l,m
4	6157875	6172549	18,19_e,f,g,o,i,l,m	4	6165163	6166460	14_e,f,o,i,l,m
4	6165163	6169225	5,6_e,f,o,i,l,m	4	6165163	6172549	18,19_e,f,o,i,l,m
4	6168352	6169225	5,6_e,f,h,t,v,o,i,l,m	4	6168352	6172549	18,19_e,f,h,t,v,o,i,l,m
4	6178445	6179571	14_u	4	6242976	6243815	5,6_e,o,i,l,m
4	6274545	6281544	17_u	4	6279019	6281544	14_u
4	6339149	6340304	17_k,s	4	6339468	6340017	9,5,6_e,o,i,l,m
4	6376254	6379405	1,16,5,6_t	4	6414050	6414635	16,5,6_e,o,i,l,m
4	6414050	6427211	16,5,6_e,o,i,l,m	4	6426613	6427211	16,5,6_e,t,o,i,l,m
4	6505631	6506263	17_p,s	4	6624860	6626153	18,19_q
4	6624860	6639548	18,19_t	4	6625288	6626153	18,19_q
4	6625288	6639548	18,19_t	4	6638801	6639548	1,9,5,6_t
4	6808233	6809890	15_n	4	6894943	6895537	9,5,6_e,o,i,l,m
4	6894943	6911515	18,19_e,o,i,l,m	4	6894943	6912406	18,19_e,o,i,l,m
4	6910438	6911515	18,19_q	4	6910438	6912406	18,19_q
4	6911512	6912406	18,19_q	4	7028409	7035097	15_n
4	7034245	7034786	16,5,6_a,e,f,g,h,t,v,o,i,l,m	4	7034288	7035097	17_n
4	7058597	7060691	14_u	4	7154773	7155369	16,5,6_e,o,i,l,m
4	7230729	7231596	9,5,6_e,o,i,l,m	4	7230729	7236655	9,5,6_q
4	7235155	7236655	18,19_q	4	7383412	7384278	9,5,6_a,e,f,g,h,t,o,i,l,m
4	7403989	7404773	18,19_q	4	7652560	7653124	5,6_e,o,i,l,m
4	7830146	7830889	5,6_c,h,t,o,i,l,m	4	7830146	7838937	5,6_s
4	7838344	7838937	17_s	4	7896286	7906363	17_a,e,g,t,o,i,l,m
4	7896286	7906947	17_p	4	7905242	7906363	17_a,e,g,t,o,i,l,m
4	7905242	7906947	17_p	4	7905767	7906363	16,5,6_a,e,g,t,o,i,l,m
4	7905767	7906947	16,5,6_p	4	7909648	7910807	1,7,8,5,6_a,f,g
4	8063572	8065698	14_a,e,f,g,h,t,v,o,i,l,m	4	8421094	8421899	17_p
4	8564177	8564943	5,6_e,f,o,i,l,m	4	8597588	8598449	18,19_q
4	8725079	8725654	16,5,6_e,o,i,l,m	4	8725079	8730452	14_e,o,i,l,m
4	8728401	8730452	14_f,g,o,i,l,m	4	8802132	8802736	14_a,e,g,k,t,o,i,l,m
4	8899075	8899962	9,5,6_e,t,o,i,l,m	4	8947216	8947757	9,5,6_e,o,i,l,m
4	8957980	8958745	16,5,6_e,f,g,o,i,l,m	4	9078980	9079478	17_a,k,s
4	9111590	9112139	16,5,6_a,e,g,t,o,i,l,m	4	9146457	9147036	5,6_o,i,l,m
4	9175159	9190300	18,19_e,o,i,l,m	4	9188732	9190300	16,5,6_e,o,i,l,m
4	9189217	9190213	5,6_a,e,g,h,k,t,o,i,l,m	4	9283790	9284437	17_k,s
4	9368738	9369364	17_s	4	9522692	9523970	1,7,8,9,16_a,f,g,o,i,l,m
4	9687723	9688999	7,8,9_a,f,g	4	9687723	9701294	12_a,f,g
4	9761907	9763826	14_n	4	9840704	9842620	14_n
4	10395360	10397395	14_a,f,g,o,i,l,m	4	10396440	10397395	14_f,o,i,l,m
4	10870794	10871696	18,19_q	4	10999811	11000857	9,5,6_a,c,e,g,k,t,o,i,l,m
4	11192286	11193580	14_a,c,e,f,g,h,k,v,o,i,l,m	4	11224171	11225123	17_p
4	11360881	11362955	14_a,e,f,g,o,i,l,m	4	11380712	11381843	5,6_e,o,i,l,m
4	11696275	11696737	17_s	4	11751734	11752263	8,5,6_h,t,o,i,l,m
4	11978732	11979613	18,19_q	4	12365074	12365615	1,9,5,6_e,f,h,t,v,o,i,l,m
4	12606487	12607056	5,6,16_o,i,l,m	4	12698510	12699371	5,6_a,e,g,t,o,i,l,m
4	12713709	12714145	22_v,o,i,l,m	4	13123878	13124462	1,16,5,6_e,u,o,i,l,m
4	13536678	13537675	1,9,5,6_a,c,e,g,h,k,t,o,i,l,m	4	14390268	14390993	14_a,e,f,g,h,t,v,o,i,l,m

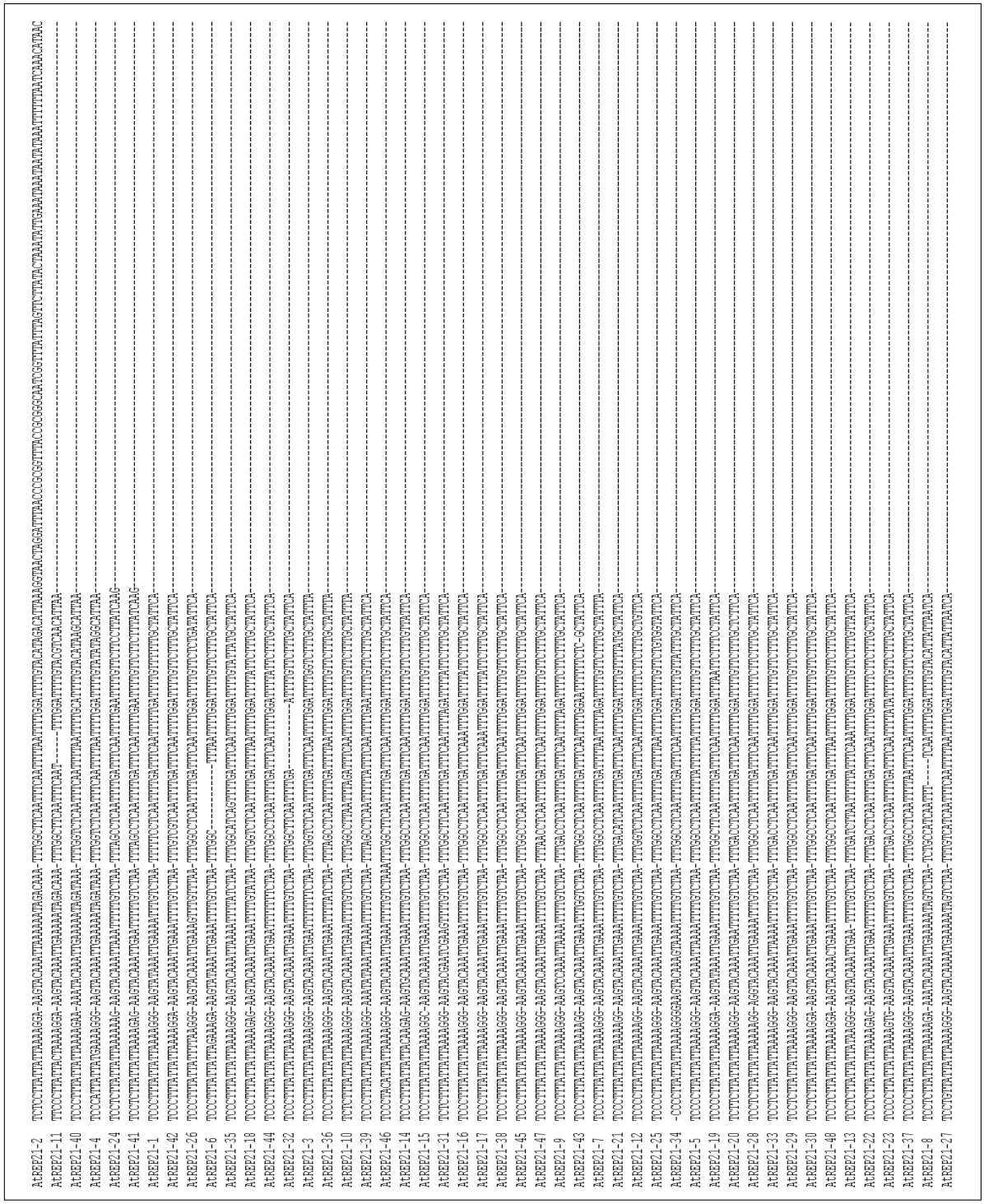
Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
4	14457717	14458289	16,5,6_e,f,g,o,i,l,m	4	14978041	14979219	17_p,s
4	15441657	15442199	5,6_e,f,h,t,v,o,i,l,m	4	15816059	15816689	17_k,s
4	16054517	16055478	18,19_q	4	16054517	16055646	18,19_q
4	16705430	16705866	22_v	4	16863268	16865074	14_a,c,e,f,g,o,i,l,m
4	18425752	18426661	17_p	5	347013	347800	18,19_q
5	347013	348081	18,19_q	5	347013	348363	18,19_q
5	347013	357010	17_q	5	349807	350407	5,6_e,o,i,l,m
5	349807	356200	5,6_u	5	353871	357010	17_p
5	354137	356200	14_u	5	1180518	1181401	18,19_q
5	1896577	1898645	14_a,f,o,i,l,m	5	1897708	1898645	14_o,i,l,m
5	2603154	2605538	1,7,16,5,6_f	5	3114303	3116667	1,7,16,5,6_f,g,o,i,l,m
5	4312439	4313476	14_o,i,l,m	5	4312439	4314868	14_a,e,f,g,h,t,v,o,i,l,m
5	4780213	4782230	14_a,e,f,g,h,t,v,o,i,l,m	5	4781293	4782230	14_f,o,i,l,m
5	4936964	4937493	5,6_o,i,l,m	5	4970262	4971027	16,5,6_a,e,f,g,h,v,o,i,l,m
5	5062375	5063617	18,19_q	5	5128495	5129836	14_u,o,i,l,m
5	5427959	5430064	14_a,e,f,g,h,t,v,o,i,l,m	5	5625061	5625622	9,5,6_e,o,i,l,m
5	6072552	6073725	18,19_q	5	6171148	6171682	8,5,6_t
5	6419829	6420379	5,6_o,i,l,m	5	6701929	6702419	16,5,6_e,o,i,l,m
5	7056140	7057392	20_r	5	7069991	7070518	16,5,6_e,o,i,l,m
5	7153186	7154440	14_a,e,f,g,h,t,v,o,i,l,m	5	7235593	7237495	23_w
5	7352036	7352767	18,19_q	5	7417334	7417910	16,5,6_e,f,h,o,i,l,m
5	7488094	7488979	9,5,6_e,o,i,l,m	5	7700641	7701183	16,5,6_a,e,f,g,h,t,v,o,i,l,m
5	7700684	7713225	17_u	5	7711134	7713225	14_u
5	7820803	7834776	18,19_q	5	7833878	7834776	18,19_q
5	8177672	8178109	22_v	5	8221101	8223983	17_p
5	8280721	8281723	9,5,6_e,f,g,o,i,l,m	5	8312119	8314484	15_n
5	8312119	8316232	18,19_n	5	8601822	8602696	5,6_a,e,g,t,u,o,i,l,m
5	8682982	8685093	14_u	5	8731241	8744668	18,19_g
5	8731241	8749472	18,19_e,o,i,l,m	5	8744628	8746703	14_a,e,f,g,h,t,v,o,i,l,m
5	8745762	8746703	14_o,i,l,m	5	8748928	8749472	5,6_e,o,i,l,m
5	8748928	8766668	5,6_o,i,l,m	5	8748928	8767805	5,6_a,f,o,i,l,m
5	8765712	8766668	14_o,i,l,m	5	8765712	8767805	14_a,f,o,i,l,m
5	8808998	8809612	22_v	5	9068174	9068737	5,6,9_o,i,l,m
5	9133956	9134726	14_a,e,f,g,o,i,l,m	5	9177237	9179290	1,7,8,9_g
5	9177237	9195851	17_g	5	9177772	9179290	1,7,8,9_g
5	9177772	9195851	17_g	5	9194778	9195851	17_p
5	9204362	9205318	18,19_q	5	9204362	9222910	18,19_p
5	9204362	9224217	18,19_f,o,i,l,m	5	9222041	9222910	17_p
5	9222041	9224217	17_f,o,i,l,m	5	9298129	9298901	5,6_e,f,v,o,i,l,m
5	9362604	9363781	17_p	5	9362604	9369931	17_h,t,o,i,l,m
5	9368947	9369931	5,6_h,t,o,i,l,m	5	9371975	9374096	15_n
5	9424372	9443934	22_q	5	9442643	9443934	18,19_q
5	9545741	9546335	1,5,6_e,o,i,l,m	5	9557662	9558218	9,16,5,6_a,e,f,g,h,t,v,o,i,l,m
5	9557662	9559509	14_a,e,f,g,h,t,v,o,i,l,m	5	9558554	9559509	14_o,i,l,m
5	9590413	9592538	14_a,e,f,g,h,o,i,l,m	5	9590413	9593677	14_q
5	9591750	9592538	14_a,e,f,g,h,o,i,l,m	5	9591750	9593677	14_q
5	9664620	9666387	21_n	5	9697539	9702726	18,19_d
5	9697539	9702901	18,19_q	5	9747581	9755210	17_q
5	9753628	9755210	18,19_q	5	9762355	9763122	1,9,5,6_a,c,e,f,g,h,t,o,i,l,m
5	9762355	9770645	9,5,6_a,c,e,f,g,h,t,o,i,l,m	5	9770086	9770645	9,5,6_t,v
5	9854350	9854894	5,6_o,i,l,m	5	9854350	9858650	5,6_e,f,o,i,l,m
5	9854350	9867709	5,6_o,i,l,m	5	9866845	9867709	5,6,9_o,i,l,m
5	9924332	9933944	23_s,w	5	9962951	9965020	14_a,e,f,g,o,i,l,m
5	9962951	9979176	1,9,5,6_a,e,f,g,o,i,l,m	5	9964081	9965020	14_f,o,i,l,m
5	9964081	9979176	1,9,5,6_f,o,i,l,m	5	9978590	9979176	1,9,5,6_e,o,i,l,m
5	10041816	10042377	16,5,6_e,f,h,t,v,o,i,l,m	5	10041816	10049098	5,6_e,f,h,t,v,o,i,l,m
5	10048060	10049098	5,6_o,i,l,m	5	10048060	10067138	21,15_o,i,l,m
5	10151181	10162511	12_j	5	10175594	10178909	3,10,13_c,e,h,k,s
5	10175594	10187464	3,10,13_d	5	10210116	10212040	1,16,5,6_a,e,f,g,v,o,i,l,m
5	10210116	10212216	1,16,5,6_a,e,f,g,t,o,i,l,m	5	10268947	10270535	18,19_p,s
5	10268947	10271036	17_p,s	5	10288010	10289730	21_n
5	10369955	10371413	17_p	5	10369955	10372171	17_p
5	10371511	10372171	17_p	5	10714428	10715625	14_a,e,f,g,h,t,v,o,i,l,m
5	10714428	10719038	14_n	5	10715625	10719038	21_n
5	10716665	10719038	15_n	5	10859898	10874660	18,19_t,o,i,l,m
5	10873665	10874660	1,9,5,6_t,o,i,l,m	5	11019346	11022155	1,7,16_a,e,f,g,h,o,i,l,m
5	11025340	11026284	14_f,o,i,l,m	5	11025340	11027899	14_a,f,g,o,i,l,m
5	11656228	11658338	17_p	5	12107520	12108954	18,19_q
5	12325051	12333318	5,6_a,c,e,f,g,t,o,i,l,m	5	12325051	12339086	5,6_a,c,e,f,g,t,o,i,l,m
5	12465850	12468832	3,10,13_h,k	5	12677481	12678504	17_p
5	12977706	12981849	3,10,13_a,c,g,h,k,t	5	13248899	13249302	5,6_a,e,g,t,o,i,l,m
5	13260985	13269716	1_a,g	5	13409670	13410772	7,8,9_g
5	13515784	13517561	21_n	5	13515784	13532331	15_n
5	13517548	13528645	2_b	5	13530139	13532331	15_n

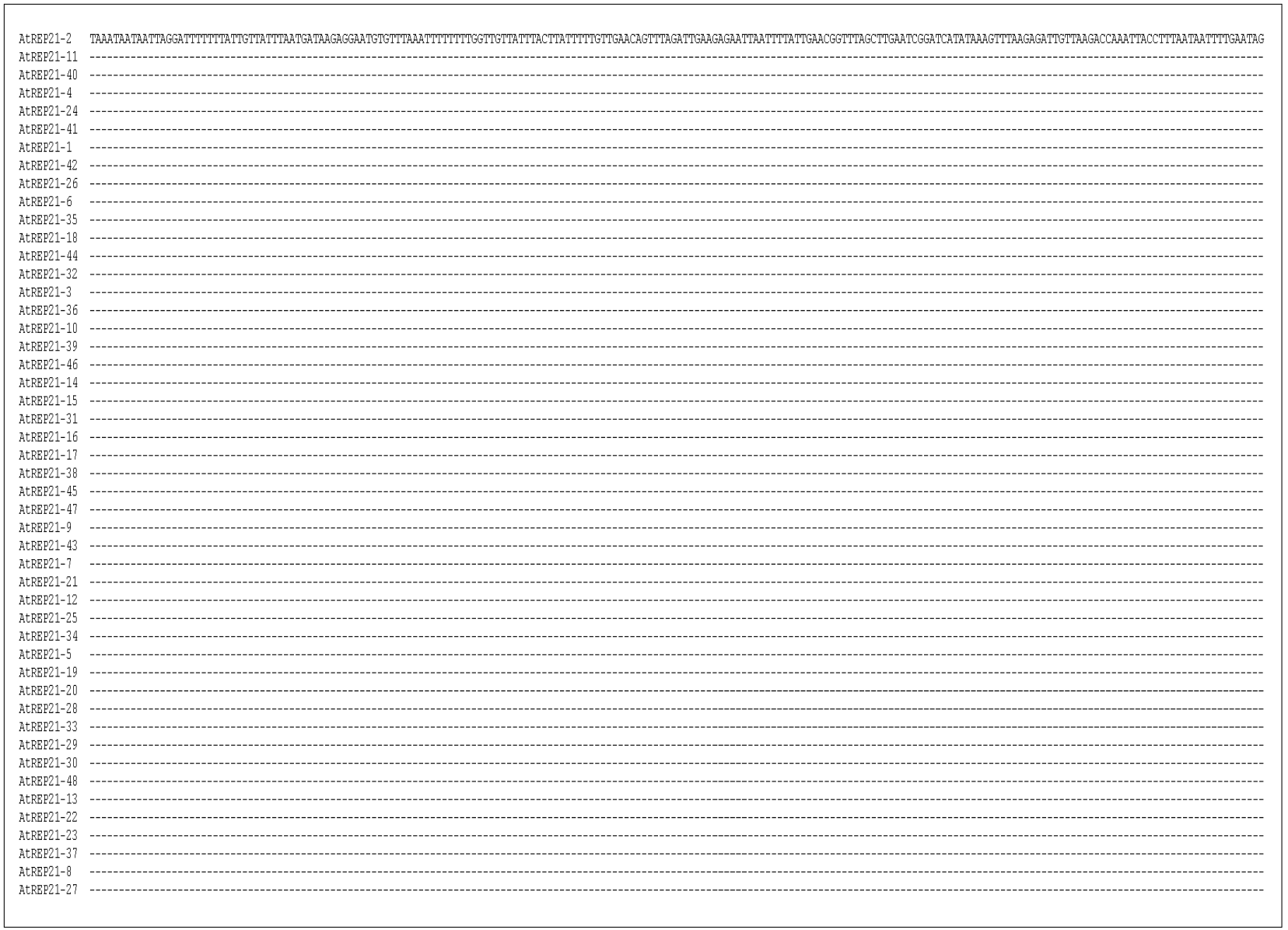
Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
5	13530139	13538063	17_n	5	13537129	13538063	17_p
5	13564241	13566471	1,16,5,6_f,g,o,i,l,m	5	13625526	13625941	22_v
5	13625526	13636887	22_f,g	5	13634527	13636660	15_n
5	13634527	13639224	16,5,6_n	5	13636847	13639224	16,5,6_a,e,f,g,v,o,i,l,m
5	13693669	13696366	18,19_q	5	13695042	13696259	1,9,5,6_e,o,i,l,m
5	13745591	13747699	15_n	5	13745591	13752563	23_n
5	13746188	13747699	15_g	5	13746188	13752563	23_g
5	13747877	13752563	23_w	5	13750260	13751723	9,5,6_e,o,i,l,m
5	13750260	13752184	9,5,6_a,c,e,g,k,t,o,i,l,m	5	13750839	13751723	7,8_e,o,i,l,m
5	13750839	13752184	7,8_a,c,e,g,k,t,o,i,l,m	5	13803768	13804567	17_p
5	13828275	13833097	18,19_t	5	13828275	13847418	18,19_q
5	13829055	13833097	5,6_t	5	13829055	13847418	5,6_q
5	13829578	13832377	1,7,8,9,16_a,f,g	5	13829578	13844878	18,19_a,f,g
5	13829578	13846514	17_a,f,g	5	13829807	13832377	1,7,8,9,16_a,e,g
5	13829807	13844878	18,19_a,e,g	5	13829807	13846514	17_a,e,g
5	13844011	13847418	18,19_q	5	13845750	13847418	18,19_q
5	13845863	13846514	17_k,p,s	5	13909154	13909915	16,5,6_p
5	13909360	13909915	16,5,6_e,o,i,l,m	5	13934523	13936019	14_u
5	13987319	14001552	20_u	5	13987319	14004678	20_a,f,g,h
5	13994282	14001552	5,6_u	5	13994282	14004678	5,6_a,f,g,h
5	13998487	14001552	17_u	5	13998487	14004678	17_a,f,g,h
5	13999365	14001552	14_u	5	13999365	14004678	14_a,f,g,h
5	14001552	14004678	3,10,13_a,f,g,h	5	14003080	14004678	1_a,f,g,h
5	14003080	14022250	1_e,f,g,v,o,i,l,m	5	14020014	14022250	1,7,16,5,6_e,f,g,v,o,i,l,m
5	14074576	14075785	7,8,9_g	5	14096653	14097531	9,5,6_a,e,f,g,h,t,v,o,i,l,m
5	14097993	14100064	14_e,f,k,u,o,i,l,m	5	14097993	14114335	15_e,f,k,u,o,i,l,m
5	14097993	14115336	18,19_e,f,k,u,o,i,l,m	5	14099127	14100064	14_o,i,l,m
5	14099127	14114335	15_o,i,l,m	5	14099127	14115336	18,19_o,i,l,m
5	14112128	14114335	15_n	5	14112128	14115336	18,19_n
5	14112128	14121437	5,6_n	5	14120897	14121437	5,6_a,e,f,g,h,t,v,o,i,l,m
5	14160833	14168523	14_n	5	14160833	14172882	15_n
5	14160924	14168523	14_a,c,e,f,g,o,i,l,m	5	14160924	14172882	15_a,c,e,f,g,o,i,l,m
5	14228077	14229396	5,6_e,o,i,l,m	5	14228077	14233583	9_e,o,i,l,m
5	14230952	14232718	21_n	5	14232832	14233583	9_c,h,t
5	14281942	14282838	5,6_e,g,o,i,l,m	5	14283665	14285185	8,9,16,5,6_f,g,o,i,l,m
5	14340623	14341650	18,19_q	5	14340623	14343838	18,19_q
5	14340623	14346037	18,19_q	5	14340623	14348225	18,19_q
5	14340623	14350412	18,19_q	5	14340623	14352607	18,19_q
5	14340623	14354804	18,19_q	5	14340623	14357001	18,19_q
5	14340623	14359196	18,19_q	5	14410741	14411738	5,6_e,g,o,i,l,m
5	14412263	14413074	5,6_v	5	14456446	14457566	7,8,9_g
5	14456446	14457722	7,8,9_a,f,g	5	14456446	14458539	7,8,9_f,g
5	14456446	14468864	7,8,9_q	5	14456446	14471037	7,8,9_e,g,o,i,l,m
5	14457682	14472373	5,6_f	5	14457709	14458539	8_f,g
5	14457709	14468864	8_q	5	14457709	14471037	8_e,g,o,i,l,m
5	14470040	14471037	5,6_e,g,o,i,l,m	5	14471562	14472373	5,6_v
5	14515746	14516866	7,8,9_g	5	14515746	14517022	7,8,9_a,f,g
5	14515746	14517839	7,8,9_f,g	5	14517009	14517839	8_f,g
5	14619921	14620983	17_p	5	14629815	14630291	18,19_q
5	14676595	14693855	4_h,k,t	5	14710667	14711320	18,19_q
5	14749423	14750032	18,19_q	5	14765366	14778412	14_e,t
5	14827501	14828054	5,6_e,f,h,v,o,i,l,m	5	14863819	14866035	16,5,6_a,e,f,g,v,o,i,l,m
5	14888548	14891282	18,19_o,i,l,m	5	14888548	14898042	18,19_e,o,i,l,m
5	14890686	14891282	5,6,9_o,i,l,m	5	14890686	14898042	9,5,6_e,o,i,l,m
5	14897446	14898042	16,5,6_e,o,i,l,m	5	14973303	14973996	17_k,s
5	14973303	14985134	18,19_k,s	5	15002026	15002461	22_v
5	15002026	15021178	17_v	5	15018417	15021178	17_b,p,s
5	15048144	15049041	9,5,6_a,e,g,t,o,i,l,m	5	15048144	15060657	14_a,e,g,t,o,i,l,m
5	15048144	15061441	7,8_a,e,g,t,o,i,l,m	5	15052402	15070918	17_q
5	15059364	15060657	14_a,c,e,f,g,h,k,v,o,i,l,m	5	15059364	15061441	7,8_a,c,e,f,g,h,k,v,o,i,l,m
5	15104963	15106475	7_o,i,l,m	5	15105230	15106475	7_a,f,g
5	15166075	15180351	12_a,e,g,t,o,i,l,m	5	15179485	15180351	9,5,6_a,e,g,t,o,i,l,m
5	15230129	15230552	22_v,o,i,l,m	5	15358444	15359925	14_u
5	15402830	15405069	1,16,5,6_a,f,g,o,i,l,m	5	15593410	15594654	14_a,e,f,o,i,l,m
5	15598037	15598570	16,5,6_e,o,i,l,m	5	15598037	15608555	16,5,6_q
5	15621899	15623830	14_a,e,f,g,h,t,v,o,i,l,m	5	15622890	15623830	14_g,o,i,l,m
5	15732174	15732902	18,19_q	5	15796272	15797049	9,5,6_a,e,f,g,h,v,o,i,l,m
5	15796272	15808961	9,5,6_s	5	15808334	15808961	17_s
5	15887181	15899993	18,19_a,e,f,g,h,t,v,o,i,l,m	5	15899521	15899993	5,6_a,e,f,g,h,t,v,o,i,l,m
5	16047777	16048614	5,6_e,f,h,t,v,o,i,l,m	5	16373703	16375771	14_a,e,f,g,o,i,l,m
5	16373703	16376499	14_a,f,g,o,i,l,m	5	16373703	16391078	14_e,t,o,i,l,m
5	16375771	16376499	14_a,f,g,o,i,l,m	5	16375771	16391078	14_e,t,o,i,l,m
5	16390227	16391078	5,6_e,t,o,i,l,m	5	16390227	16405234	5,6_v,o,i,l,m
5	16404919	16405234	22_v,o,i,l,m	5	16730079	16730620	16,5,6_e,o,i,l,m

Chr	Début	Fin	Combinaisons 5'_3'	Chr	Début	Fin	Combinaisons 5'_3'
5	16888151	16888706	16,5,6_e,t,o,i,l,m	5	17058985	17060057	17_p
5	17058985	17061128	17_e,f,h,t,v,o,i,l,m	5	17060586	17061128	5,6_e,f,h,t,v,o,i,l,m
5	17292371	17293226	18,19_q	5	17376551	17377328	17_p
5	17382253	17383210	14_o,i,l,m	5	17382253	17384350	14_a,f,g,h,o,i,l,m
5	17495893	17499273	18,19_e,o,i,l,m	5	17498379	17499273	5,6_e,o,i,l,m
5	17595705	17596808	17_p,s	5	17743649	17744211	16,5,6_a,e,f,g,h,t,v,o,i,l,m
5	17792956	17795028	14_a,e,f,g,u,o,i,l,m	5	17813365	17814011	14_a,e,f,g,h,t,v,o,i,l,m
5	17813365	17830985	2_a,e,f,g,h,t,v,o,i,l,m	5	17902823	17903177	5,6_e,f,h,t,v,o,i,l,m
5	17972782	17974401	14_u	5	18130869	18131229	22_v
5	18164780	18165133	22_v	5	18571534	18573046	18,19_p
5	18571534	18573689	18,19_q	5	18572135	18573046	17_p
5	18572135	18573689	17_q	5	18936820	18937449	17_k,s
5	19035847	19036792	17_p	5	19097775	19098655	18,19_q
5	19118785	19119334	9,5,6_a,e,g,t,o,i,l,m	5	19169645	19170889	11_a,e,g,t,o,i,l,m
5	19169645	19170972	5,6_a,e,g,t,o,i,l,m	5	19425264	19426582	9_a,c,e,f,g,h,k,t,o,i,l,m
5	19425604	19426582	9_o,i,l,m	5	19430288	19438116	5,6_a,f,g,o,i,l,m
5	19436174	19438116	1,7,16,5,6_a,f,g,o,i,l,m	5	19559946	19562125	22_h,v
5	19757741	19759842	14_a,c,e,f,g,t,u,o,i,l,m	5	19929928	19930772	9,5,6_e,t,o,i,l,m
5	19947437	19948783	14_u	5	20295840	20297872	14_a,e,g,t,o,i,l,m
5	20295840	20299209	14_a,e,g,t,o,i,l,m	5	20298271	20299209	14_o,i,l,m
5	20691460	20692047	16,5,6_e,o,i,l,m	5	20691460	20699631	14_e,o,i,l,m
5	20697542	20699631	14_a,f,g,o,i,l,m	5	20698679	20699631	14_f,o,i,l,m
5	21033061	21033794	9,5,6_t	5	21117317	21118061	15_n
5	21547315	21549133	14_a,e,f,g,o,i,l,m	5	21793013	21794257	14_a,e,f,g,h,t,v,o,i,l,m
5	21793013	21795748	3,10,13_a,e,f,g,h,t,v,o,i,l,m	5	22038199	22039085	9,5,6_e,o,i,l,m
5	22248178	22249489	14_a,f,g,o,i,l,m	5	22480884	22481427	5,6_e,f,v,o,i,l,m
5	22638404	22638965	5,6_a,e,f,g,h,t,u,v,o,i,l,m	5	23067189	23067916	5,6_t,o,i,l,m
5	23146567	23147439	5,6_e,o,i,l,m	5	23448769	23450110	20_r
5	23679230	23679805	9,16,5,6_e,o,i,l,m	5	24019309	24020554	20_r
5	24117932	24135775	18,19_e,f,h,o,i,l,m	5	24135186	24135775	16,5,6_e,f,h,o,i,l,m
5	24330867	24331720	5,6_a,e,g,v,o,i,l,m	5	24361316	24362769	18,19_q
5	24452937	24453833	18,19_q	5	24466038	24466600	5,6_a,e,g,t,o,i,l,m
5	24492517	24493816	14_a,e,f,g,h,o,i,l,m	5	24603857	24604281	5,6_a,e,g,t,o,i,l,m
5	25196003	25196605	5,6_e,o,i,l,m	5	25281737	25282562	14_a,e,f,g,h,t,v,o,i,l,m
5	25861822	25862589	5,6_a,e,f,g,h,o,i,l,m	5	25921958	25922840	14_a,f,g,h,o,i,l,m
5	25953125	25953686	17_p	5	26250342	26250757	22_v
5	26500247	26502630	16_a,f,g,o,i,l,m	5	26804332	26804879	5,6_e,o,i,l,m
5	26908864	26910624	5,6_e,t,o,i,l,m				

5.2 Annexe 2 : Alignement multiple de la famille AtREP21

L'alignement multiple des 48 copies d'AtREP21 a été réalisé avec ClustalW 1.83, les paramètre d'ouverture d'un gap était de 25 et celui de l'extension d'un gap était de 0,01.




AtREP21-2 TAAATAATAATTAGGATTTTTTTATTGTTATTTAATGATAAGAGGAATGTGTTTAAATTTTTTTTGGTTGTTATTTACTTATTTTTGTTGAACAGTTTAGATTGAAGAGAATTAATTTTATTGAACGGTTTAGCTTGAATCGGATCATATAAAGTTTAAGAGATTGTTAAGACCAAATTACCTTTAATAATTTTGAATAG
AtREP21-11
AtREP21-40
AtREP21-4
AtREP21-24
AtREP21-41
AtREP21-1
AtREP21-42
AtREP21-26
AtREP21-6
AtREP21-35
AtREP21-18
AtREP21-44
AtREP21-32
AtREP21-3
AtREP21-36
AtREP21-10
AtREP21-39
AtREP21-46
AtREP21-14
AtREP21-15
AtREP21-31
AtREP21-16
AtREP21-17
AtREP21-38
AtREP21-45
AtREP21-47
AtREP21-9
AtREP21-43
AtREP21-7
AtREP21-21
AtREP21-12
AtREP21-25
AtREP21-34
AtREP21-5
AtREP21-19
AtREP21-20
AtREP21-28
AtREP21-33
AtREP21-29
AtREP21-30
AtREP21-48
AtREP21-13
AtREP21-22
AtREP21-23
AtREP21-37
AtREP21-8
AtREP21-27

```
AtREP21-2   CAAAGGAGAAAAATCCAAAATTGAAATCAAAATTGAGGCCAAATTAGACAAATTTTCAATTTGTACCTCCTTTTTTAAATAATAAGAGGATGATTTGGTCAACTCAATCTATTTAAGCTTATAGGATCCACGTTATTAATTT-CGGGTGGGTTAAAAGTAATG----------
AtREP21-11  ----------------------------------------------------------GGATAATTTGGGTTAACTTATCTATTAAAGCTTATAGGATCTGCATTTTTAATTT-TGGGTGGGTTTGGAGTAATG----------
AtREP21-40  ----------------------------------------------------------GGGTAATTTGGGTTAACTCAATTTATTAAAGCTTATAGGATCTACATTTTTAATTT-TGGGTGGGTTAGGAGTAATG----------
AtREP21-4   ----------------------------------------------------------GGGTAATTTGGGTCAACTCAATCTATTAAAGCTTATAGGATCCACGTTTTTTAATTT-CGGGTGGGTTAGGAGTAATG----------
AtREP21-24  ----------GGTAAAAATCGTCCAAACCAAAATACTAAATTAAACAGGATCCGGTGAGTTATGTTTTAATG----------GGTAAAATCGTCCAAACCAAAATACTAAATTAAACAGGATCCGGTGAGTTATGTTTTAATG----------
AtREP21-41  ----------GGTAAAATCGTCCAAACCAAATACTAAATTAAACAGGATCCGCGTTTTACATAGACGGGTGAGTTATGTTTTAATG----------GGTAAAATCGTCCAAACCAAATACTAAATTAAACAGGATCCGCGTTTTACATAGACGGGTGAGTTATGTTTTAATG----------
AtREP21-1   ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-42  ----------AAATCATTAT----------GGGTAATTTGGTCTTAACA----------AAATCATTAT----------GGGTAATTTGGTCTTAACA----------
AtREP21-26  ----------AAATCATTAA----------GGGTAATTTGCTATTAACA----------AAATCATTAA----------GGGTAATTTGCTATTAACA----------
AtREP21-6   ----------AAATTATTAA----------GGGTAATTTGGTCTTAACA----------AAATTATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-35  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-18  ----------AAATTTATTAA----------GGGTAGTTTGCTCTTAACA----------AAATTTATTAA----------GGGTAGTTTGCTCTTAACA----------
AtREP21-44  ----------AAATTATTAA----------GGCTAATTTGGTCTTAACA----------AAATTATTAA----------GGCTAATTTGGTCTTAACA----------
AtREP21-32  ----------AAATCAGTAA----------AGGTAATTTGGTCTTAACA----------AAATCAGTAA----------AGGTAATTTGGTCTTAACA----------
AtREP21-3   ----------AAATCATTAA----------GGGCAATTTGCTATTAACA----------AAATCATTAA----------GGGCAATTTGCTATTAACA----------
AtREP21-36  ----------AAATCATTAA----------GGGTAATTTGTTCTTAACA----------AAATCATTAA----------GGGTAATTTGTTCTTAACA----------
AtREP21-10  ----------AAATCATTAA----------GGATAATTTGGTCTTAACA----------AAATCATTAA----------GGATAATTTGGTCTTAACA----------
AtREP21-39  ----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------
AtREP21-46  ----------AAATCATTAA----------GAGTAATTTGCTATTAACA----------AAATCATTAA----------GAGTAATTTGCTATTAACA----------
AtREP21-14  ----------AAATCATTAA----------GGATAATTTGGTCTTAACA----------AAATCATTAA----------GGATAATTTGGTCTTAACA----------
AtREP21-15  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-31  ----------AAATCATTAA----------GGGTAAATTGGTCTTAACA----------AAATCATTAA----------GGGTAAATTGGTCTTAACA----------
AtREP21-16  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-17  ----------AAATTATTAA----------GGGTAATTTGCTCTTAACA----------AAATTATTAA----------GGGTAATTTGCTCTTAACA----------
AtREP21-38  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-45  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-47  ----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------
AtREP21-9   ----------AAATCATTAA----------GGATAATTTGGTCTTAACA----------AAATCATTAA----------GGATAATTTGGTCTTAACA----------
AtREP21-43  ----------AAATCATTTA----------GGATAATTTGGTCTTAACA----------AAATCATTTA----------GGATAATTTGGTCTTAACA----------
AtREP21-7   ----------AAATCATTAA----------GGATAATTTGCTCTTAACA----------AAATCATTAA----------GGATAATTTGCTCTTAACA----------
AtREP21-21  ----------AAATTATTAA----------GGGTAATTTGGTCTTAACA----------AAATTATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-12  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-25  ----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------
AtREP21-34  ----------AAATCATTAA----------AGATAATTTGGTCTTAATA----------AAATCATTAA----------AGATAATTTGGTCTTAATA----------
AtREP21-5   ----------AAATCATTAA----------GGATAATTTGGTCTTAACA----------AAATCATTAA----------GGATAATTTGGTCTTAACA----------
AtREP21-19  ----------AAAACATTAA----------GGGTAATTTGGTCTTAACA----------AAAACATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-20  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-28  ----------AAATTATTAA----------AGGTAATTTGGTCTTAACA----------AAATTATTAA----------AGGTAATTTGGTCTTAACA----------
AtREP21-33  ----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------
AtREP21-29  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-30  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-48  ----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------AAATCATTAA----------GGGTAATTTGCTCTTAACA----------
AtREP21-13  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-22  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-23  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-37  ----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------AAATCATTAA----------GGGTAATTTGGTCTTAACA----------
AtREP21-8   ----------AAAATCGTCTA----------TGGTAATTTGGTCATCACTCTATATTCCTTAAATACACTATATTTCAAATTTAAATAAT----------
AtREP21-27  ----------TAATCGTCAA----------GGGTAATTTGCTCATCGCTCTATATTCCTTAAACACACTATATTTCAACTTTAAATAAT----------
```

```
AtREP21-2  ----------TAAATATGAGTATATTCCAT----------CCCAAAAAGAACTATAATCACAAAAA----------
AtREP21-11 ----------TAAATATGAGCATATTTCAT----------CCCAAAGAGAATTATAATCACAAAAA----------
AtREP21-40 ----------TATATATGAGCATATTCCA----------CTCAAAAAGAATTATAATCACAAAAA----------
AtREP21-4  ----------TAAATATGAGCATATTCCAT----------CCCAAAAAGAATTATAATCACAAAAA----------
AtREP21-24 ATTCACCCTCCTCACTAATCACTAATAAATGACATAGACGAGCTTATTCGATTCTTCTAACTACCATAATCATTCACTAATAAATTTATTCTCTTTAAATCTAAACATAATTTCACCACTTATCTAAATTATAGCCAATCATTATGAACTAAACCA
AtREP21-41 ATTCACCCTCCTCACTAATCACTATAAAATCGGCATAGACGAGCTTATTCGATTCTTCTAACTACCATAATCATTCACTAATAAATTTATTCTCTTTAAATCTAACCATAATCACTAAACATTATTTCACCACTTATCTAAATTATAGCCAATCATTATGAACTAAACCA
AtREP21-1  ----------AATCTTAAACCGTTCAA----------
AtREP21-42 ----------AATATAAACCGTTCAA----------
AtREP21-26 ----------AATATAAACCATTCAA----------
AtREP21-6  ----------AATCTTAAACCGTTCAA----------
AtREP21-35 ----------AATCTTAAACCGTTCAA----------
AtREP21-18 ----------AATCTTAGCTATTCAA----------
AtREP21-44 ----------AATCTTAAACCGTTCAA----------
AtREP21-32 ----------AATCTTAAACCGTTCAA----------
AtREP21-3  ----------
AtREP21-36 ----------AATCTTAAACCGTTCAATAAAATTAATTCCCTTCAATCTTAAACCGCTCAA----------
AtREP21-10 ----------AATCTTAAAATGTTCAA----------
AtREP21-39 ----------AATCTTAAATCGTTCAA----------
AtREP21-46 ----------AATCTTAAACCGTTTAA----------
AtREP21-14 ----------AATCTTAAATCGTTCAA----------
AtREP21-15 ----------AATCTTAAACCGTTCAA----------
AtREP21-31 ----------AATCTTAAACCGTTAAA----------
AtREP21-16 ----------AATCTTAAACCGTTCAA----------
AtREP21-17 ----------AATCTTAAACCGTTCAA----------
AtREP21-38 ----------AATCTTAAACCGTTCAA----------
AtREP21-45 ----------AATCTTAAACCGTTCAA----------
AtREP21-47 ----------AATCTTAAACCGTTCAA----------
AtREP21-9  ----------AATCTTAAACCGTTCAA----------
AtREP21-43 ----------AATCTTAAACCGTTCAA----------
AtREP21-7  ----------AATCTTAAACCGTTCAA----------
AtREP21-21 ----------
AtREP21-12 ----------AATCTTAAACCGTTCAA----------
AtREP21-25 ----------AATCTTAAACCGTTCAA----------
AtREP21-34 ----------AATCTTAAACCGTTCAA----------
AtREP21-5  ----------AATCTTAAACCGTTCAA----------
AtREP21-19 ----------AATCTTAAACCGTTCAA----------
AtREP21-20 ----------AATCTTAAACCGTTCAA----------
AtREP21-28 ----------AATCTTAAACTGTTCAA----------
AtREP21-33 ----------AATCTTAAACCGTTCAA----------
AtREP21-29 ----------AATCTTAATCCGTTCAA----------
AtREP21-30 ----------AATCTTAAACCGTTTAA----------
AtREP21-48 ----------AATCTTAAACCGTTCAA----------
AtREP21-13 ----------AATCTTAAACCGTTCAA----------
AtREP21-22 ----------AATCTTAAACCGTTCAA----------
AtREP21-23 ----------AATCTTAAGTTATTATT----------
AtREP21-37 ----------AATCTTAAATTATTATT----------
AtREP21-8  ----------ATTAAAAAATATTATG----------
AtREP21-27 ----------ATTAAAAAATATTATGAAAAAAATATTATT----------
```


AtREP21-2 ----------A-TATTTCAAAAG----------
AtREP21-11 ----------AATATTTCAAAAAAA----------
AtREP21-40 ----------AATATTTCAAAAAA----------
AtREP21-4 ----------AATATTTCAAAAAA----------
AtREP21-24 AATTCGAATCATTAGTGAGGATCCAATCTTTGATTTGCAACTTATCCTGCAAGATATTTTCGAATATTTTTGAAAATATTAGATCCAACAACCCTAGCCTCCTTCATCACTATAAATATCAAAGTCATTCTCATCACAAACCAAATTGAGATCAAGAACGATTCGATGGTTAT-ATCATAATCAATTCATGATAGTGAAT
AtREP21-41 AATTCGAATCATTAGTGAGGATCCAATCTTTGATTTGCAACTTATCCTGCAAAATATTTTCGAATATTTTTGAAAATATTAGATCCAACAACCCTAGCCTCCTTCATCACTATAAATATCAAAGTCATTCTCATCACAAACCAAATTGAGATCAAGAACGATTCGATGGTTATTATCATAATCAATTCATGATAGTGAAT
AtREP21-1 ----------CAAAAATA----------
AtREP21-42 ----------CAAAAATA----------
AtREP21-26 ----------CAAAAGTA----------
AtREP21-6 ----------CAAAAATA----------
AtREP21-35 ----------CAAAAATA----------
AtREP21-18 ----------CAAAAATA----------
AtREP21-44 ----------CAAAAATA----------
AtREP21-32 ----------CAAAAATA----------
AtREP21-3 ----------
AtREP21-36 ----------CAAAAATA----------
AtREP21-10 ----------CAAAAATA----------
AtREP21-39 ----------CAAAAATA----------
AtREP21-46 ----------CAAAA-TA----------
AtREP21-14 ----------CAAAAATA----------
AtREP21-15 ----------CAAAAATA----------
AtREP21-31 ----------CAAAAATA----------
AtREP21-16 ----------CAAAAATA----------
AtREP21-17 ----------CAAAAAGA----------
AtREP21-38 ----------CAAAAATA----------
AtREP21-45 ----------CAAAAATA----------
AtREP21-47 ----------CAAAAATA----------
AtREP21-9 ----------CAAAAATA----------
AtREP21-43 ----------CAAAAATA----------
AtREP21-7 ----------
AtREP21-21 ----------
AtREP21-12 ----------CAAAAATA----------
AtREP21-25 ----------CAAAAATA----------
AtREP21-34 ----------CAAAAATA----------
AtREP21-5 ----------CAAAAATA----------
AtREP21-19 ----------CAAAAATA----------
AtREP21-20 ----------CAAAAATA----------
AtREP21-28 ----------CAAAAATA----------
AtREP21-33 ----------CAAAAATA----------
AtREP21-29 ----------CAAAAATA----------
AtREP21-30 ----------CAAAAATA----------
AtREP21-48 ----------CAAAAATA----------
AtREP21-13 ----------CAAAAATA----------
AtREP21-22 ----------CAAAAATA----------
AtREP21-23 ----------
AtREP21-37 ----------
AtREP21-8 ----------TTGGCAAACAATATTAAGAGCTCGCGAATT----------
AtREP21-27 ----------GTTGGTAAACAATATTAAGAGCTCGCGAATC----------

```
AtREP21-2  ----------------------------------------------------------------------------------------------------
AtREP21-11 ----------------------------------------------------------------------------------------------------
AtREP21-40 ----------------------------------------------------------------------------------------------------
AtREP21-4  ----------------------------------------------------------------------------------------------------
AtREP21-24 GATTGTTTGCATAGGTTATTCATGACGAAGTGGATGTTCTTCTTTGATGTTGATGTTCCGGTGAAAGAGACGATTTCGTCTGTGATGGAATCTTTGGAGGAATATTTTTCAGTTTACTTATTGAACTGATGCCTTATGAGTAGTCGGGTTTGTTTATTCCGGTGATGGCCCGGCCAGAACTTACAGGTAACAT
AtREP21-41 GATTGTTTGCATAGGTTATTCATGACGAAGTGGATGTTCTTCTTTGATGTTGATGTTCCGGTGAAAGAGACGATTTCGTCTGTGATGGAATCTTTGGAGGAATATTTTTCAGTTTACTTATTGAACTGATGCCTTATGAGTAGTCGGGTTTGTTTATTCCGGTGATGGCCCGGCCAGAACTTACAGGTAACAT
AtREP21-1  ----------------------------------------------------------------------------------------------------
AtREP21-42 ----------------------------------------------------------------------------------------------------
AtREP21-26 ----------------------------------------------------------------------------------------------------
AtREP21-6  ----------------------------------------------------------------------------------------------------
AtREP21-35 ----------------------------------------------------------------------------------------------------
AtREP21-18 ----------------------------------------------------------------------------------------------------
AtREP21-44 ----------------------------------------------------------------------------------------------------
AtREP21-32 ----------------------------------------------------------------------------------------------------
AtREP21-3  ----------------------------------------------------------------------------------------------------
AtREP21-36 ----------------------------------------------------------------------------------------------------
AtREP21-10 ----------------------------------------------------------------------------------------------------
AtREP21-39 ----------------------------------------------------------------------------------------------------
AtREP21-46 ----------------------------------------------------------------------------------------------------
AtREP21-14 ----------------------------------------------------------------------------------------------------
AtREP21-15 ----------------------------------------------------------------------------------------------------
AtREP21-31 ----------------------------------------------------------------------------------------------------
AtREP21-16 ----------------------------------------------------------------------------------------------------
AtREP21-17 ----------------------------------------------------------------------------------------------------
AtREP21-38 ----------------------------------------------------------------------------------------------------
AtREP21-45 ----------------------------------------------------------------------------------------------------
AtREP21-47 ----------------------------------------------------------------------------------------------------
AtREP21-9  ----------------------------------------------------------------------------------------------------
AtREP21-43 ----------------------------------------------------------------------------------------------------
AtREP21-7  ----------------------------------------------------------------------------------------------------
AtREP21-21 ----------------------------------------------------------------------------------------------------
AtREP21-12 ----------------------------------------------------------------------------------------------------
AtREP21-25 ----------------------------------------------------------------------------------------------------
AtREP21-34 ----------------------------------------------------------------------------------------------------
AtREP21-5  ----------------------------------------------------------------------------------------------------
AtREP21-19 ----------------------------------------------------------------------------------------------------
AtREP21-20 ----------------------------------------------------------------------------------------------------
AtREP21-28 ----------------------------------------------------------------------------------------------------
AtREP21-33 ----------------------------------------------------------------------------------------------------
AtREP21-29 ----------------------------------------------------------------------------------------------------
AtREP21-30 ----------------------------------------------------------------------------------------------------
AtREP21-48 ----------------------------------------------------------------------------------------------------
AtREP21-13 ----------------------------------------------------------------------------------------------------
AtREP21-22 ----------------------------------------------------------------------------------------------------
AtREP21-23 ----------------------------------------------------------------------------------------------------
AtREP21-37 ----------------------------------------------------------------------------------------------------
AtREP21-8  ----------------------------------------------------------------------------------------------------
AtREP21-27 ----------------------------------------------------------------------------------------------------
```

```
AtREP21-2  ----------------------------------------CACTTATAAAT----CGCAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-11 ----------------------------------------CACTTATAAT-----CACAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-40 ----------------------------------------CACTTATAAT-----CACAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-4  ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-24 TCATTAACTTTGTTTGATTTTTAATGATTGAACGTTAACCAAAGCCCATGTTGTTTATTTGAGTGACTTAGAATTATTAGTCACAAGTTTTTTTGAAGGTGACATATTATCAATTAAGCATCTGATTCTGTCAACACTTGGCCTGTAAAAGTCTTTTGTGCAATCCTGTTATTAGTCTCCTATGTCGGTCACGATCTTTGCTTTTA
AtREP21-41 TCATTAACTTTGTTTGATTTTTAATGATTGAACGTTAACCAAAGCCCATGTTGTTTATTTGAGTGACTTAGAATTATTAGTCACAAGTTTTTTTGAAGGTGACATATTATCAATTAAGCATCTGATTCTGTCAACACTTGGCCTGTAAAAGTCTTTTGTGCAATCCTGTTATTAGTCTCCTATGTCGGTCACGATCTTTGCTTTTA
AtREP21-1  ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-42 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-26 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-6  ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-35 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-18 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-44 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-32 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-3  ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-36 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-10 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-39 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-46 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-14 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-15 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-31 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-16 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-17 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-38 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-45 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-47 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-9  ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-43 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-7  ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-21 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-12 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-25 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-34 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-5  ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-19 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-20 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-28 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-33 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-29 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-30 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-48 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-13 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-22 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-23 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-37 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-8  ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
AtREP21-27 ----------------------------------------CACTTTTAGTATTTTTACATAA-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

```
AtREP21-2  ------------------------------------------------------------------------------------------------------------------------
AtREP21-11 ------------------------------------------------------------------------------------------------------------------------
AtREP21-40 ------------------------------------------------------------------------------------------------------------------------
AtREP21-4  ------------------------------------------------------------------------------------------------------------------------
AtREP21-24 TTGAAACATCACATGATATTAGTGAGAAAATGTTGTAAAAGATTATTTCTTTTGTTTGTACAAAATATTGAATTGTAAATTGGTTTGTATATTTCTGTTTTTTAGGAAGGAACATATCGGTTATTGATTGGTTTAGTTCAGGGACTAATCTAAATGGTATTTTTGGTTTTGTGTTTCAGAAAATGGGGAATTGAGAAACTGGCA
AtREP21-41 TTGAAATATCACATGATATTAGTGAGAAAATGTTGTAAAAGATTATTTCTTTTGTTTGTACCAAAATATTGAATTGTAAATTGGTTTGTATATTTCTGTTTTTTAGGAGGGAACATATCGGTTATTGATTGGTTTAGTTCAGGGACTAATCTAAATGGTATTTTTGGTTTTGTGTTTCAGAAAATGGGGAATTGAGAAACTGGCA
AtREP21-1  ------------------------------------------------------------------------------------------------------------------------
AtREP21-42 ------------------------------------------------------------------------------------------------------------------------
AtREP21-26 ------------------------------------------------------------------------------------------------------------------------
AtREP21-6  ------------------------------------------------------------------------------------------------------------------------
AtREP21-35 ------------------------------------------------------------------------------------------------------------------------
AtREP21-18 ------------------------------------------------------------------------------------------------------------------------
AtREP21-44 ------------------------------------------------------------------------------------------------------------------------
AtREP21-32 ------------------------------------------------------------------------------------------------------------------------
AtREP21-3  ------------------------------------------------------------------------------------------------------------------------
AtREP21-36 ------------------------------------------------------------------------------------------------------------------------
AtREP21-10 ------------------------------------------------------------------------------------------------------------------------
AtREP21-39 ------------------------------------------------------------------------------------------------------------------------
AtREP21-46 ------------------------------------------------------------------------------------------------------------------------
AtREP21-14 ------------------------------------------------------------------------------------------------------------------------
AtREP21-15 ------------------------------------------------------------------------------------------------------------------------
AtREP21-31 ------------------------------------------------------------------------------------------------------------------------
AtREP21-16 ------------------------------------------------------------------------------------------------------------------------
AtREP21-17 ------------------------------------------------------------------------------------------------------------------------
AtREP21-38 ------------------------------------------------------------------------------------------------------------------------
AtREP21-45 ------------------------------------------------------------------------------------------------------------------------
AtREP21-47 ------------------------------------------------------------------------------------------------------------------------
AtREP21-9  ------------------------------------------------------------------------------------------------------------------------
AtREP21-43 ------------------------------------------------------------------------------------------------------------------------
AtREP21-7  ------------------------------------------------------------------------------------------------------------------------
AtREP21-21 ------------------------------------------------------------------------------------------------------------------------
AtREP21-12 ------------------------------------------------------------------------------------------------------------------------
AtREP21-25 ------------------------------------------------------------------------------------------------------------------------
AtREP21-34 ------------------------------------------------------------------------------------------------------------------------
AtREP21-5  ------------------------------------------------------------------------------------------------------------------------
AtREP21-19 ------------------------------------------------------------------------------------------------------------------------
AtREP21-20 ------------------------------------------------------------------------------------------------------------------------
AtREP21-28 ------------------------------------------------------------------------------------------------------------------------
AtREP21-33 ------------------------------------------------------------------------------------------------------------------------
AtREP21-29 ------------------------------------------------------------------------------------------------------------------------
AtREP21-30 ------------------------------------------------------------------------------------------------------------------------
AtREP21-48 ------------------------------------------------------------------------------------------------------------------------
AtREP21-13 ------------------------------------------------------------------------------------------------------------------------
AtREP21-22 ------------------------------------------------------------------------------------------------------------------------
AtREP21-23 ------------------------------------------------------------------------------------------------------------------------
AtREP21-37 ------------------------------------------------------------------------------------------------------------------------
AtREP21-8  ------------------------------------------------------------------------------------------------------------------------
AtREP21-27 ------------------------------------------------------------------------------------------------------------------------
```

```
AtERF21-2  ------------------------------------------------------------------------------------------------------------
AtERF21-11 ------------------------------------------------------------------------------------------------------------
AtERF21-40 ------------------------------------------------------------------------------------------------------------
AtERF21-4  ------------------------------------------------------------------------------------------------------------
AtERF21-24 TATAGAGACGTCAAGAAGGTTCCAAGACGGGATCCGGAATGTATTCGCGATAGGCGTACACATGTCTCAGAATCTACAAGGGAAGACATAGCACAAGGCTGAGAGCAAAATCCGTTACTTCCATTACCATGGTTCGATTCTTCAGCGCCGCGAGCACGCACTCCTTATGTGCTAGACACT
AtERF21-41 TATAGACGTCAAGAAGGTTCCAAGACGGGATCCGGAATGTATTCGCGATAGGCGTACACATGTCTCAGAATCTACAAGGGAAGACATAGCACAAGGCTGAGAGCAAAATCCGTTACTTCCATTACCATGGTTCGATCTCTCAGCGCCGCGAGCACGCACTCCTTATGTGCTAGACACT
AtERF21-1  ------------------------------------------------------------------------------------------------------------
AtERF21-42 ------------------------------------------------------------------------------------------------------------
AtERF21-26 ------------------------------------------------------------------------------------------------------------
AtERF21-6  ------------------------------------------------------------------------------------------------------------
AtERF21-35 ------------------------------------------------------------------------------------------------------------
AtERF21-18 ------------------------------------------------------------------------------------------------------------
AtERF21-44 ------------------------------------------------------------------------------------------------------------
AtERF21-32 ------------------------------------------------------------------------------------------------------------
AtERF21-3  ------------------------------------------------------------------------------------------------------------
AtERF21-36 ------------------------------------------------------------------------------------------------------------
AtERF21-10 ------------------------------------------------------------------------------------------------------------
AtERF21-39 ------------------------------------------------------------------------------------------------------------
AtERF21-46 ------------------------------------------------------------------------------------------------------------
AtERF21-14 ------------------------------------------------------------------------------------------------------------
AtERF21-15 ------------------------------------------------------------------------------------------------------------
AtERF21-31 ------------------------------------------------------------------------------------------------------------
AtERF21-16 ------------------------------------------------------------------------------------------------------------
AtERF21-17 ------------------------------------------------------------------------------------------------------------
AtERF21-38 ------------------------------------------------------------------------------------------------------------
AtERF21-45 ------------------------------------------------------------------------------------------------------------
AtERF21-47 ------------------------------------------------------------------------------------------------------------
AtERF21-9  ------------------------------------------------------------------------------------------------------------
AtERF21-43 ------------------------------------------------------------------------------------------------------------
AtERF21-7  ------------------------------------------------------------------------------------------------------------
AtERF21-21 ------------------------------------------------------------------------------------------------------------
AtERF21-12 ------------------------------------------------------------------------------------------------------------
AtERF21-25 ------------------------------------------------------------------------------------------------------------
AtERF21-34 ------------------------------------------------------------------------------------------------------------
AtERF21-5  ------------------------------------------------------------------------------------------------------------
AtERF21-19 ------------------------------------------------------------------------------------------------------------
AtERF21-20 ------------------------------------------------------------------------------------------------------------
AtERF21-28 ------------------------------------------------------------------------------------------------------------
AtERF21-33 ------------------------------------------------------------------------------------------------------------
AtERF21-29 ------------------------------------------------------------------------------------------------------------
AtERF21-30 ------------------------------------------------------------------------------------------------------------
AtERF21-48 ------------------------------------------------------------------------------------------------------------
AtERF21-13 ------------------------------------------------------------------------------------------------------------
AtERF21-22 ------------------------------------------------------------------------------------------------------------
AtERF21-23 ------------------------------------------------------------------------------------------------------------
AtERF21-37 ------------------------------------------------------------------------------------------------------------
AtERF21-8  ------------------------------------------------------------------------------------------------------------
AtERF21-27 ------------------------------------------------------------------------------------------------------------
```

AtREP21-2
AtREP21-11
AtREP21-40
AtREP21-4
AtREP21-24
AtREP21-41
AtREP21-1
AtREP21-42
AtREP21-26
AtREP21-6
AtREP21-35
AtREP21-18
AtREP21-44
AtREP21-32
AtREP21-3
AtREP21-36
AtREP21-10
AtREP21-39
AtREP21-46
AtREP21-14
AtREP21-15
AtREP21-31
AtREP21-16
AtREP21-17
AtREP21-38
AtREP21-45
AtREP21-47
AtREP21-9
AtREP21-43
AtREP21-7
AtREP21-21
AtREP21-12
AtREP21-25
AtREP21-34
AtREP21-5
AtREP21-19
AtREP21-20
AtREP21-28
AtREP21-33
AtREP21-29
AtREP21-30
AtREP21-48
AtREP21-13
AtREP21-22
AtREP21-23
AtREP21-37
AtREP21-8
AtREP21-27

AcREP21-2
AcREP21-11
AcREP21-40
AcREP21-4
AcREP21-24
AcREP21-41
AcREP21-1
AcREP21-42
AcREP21-26
AcREP21-6
AcREP21-35
AcREP21-18
AcREP21-44
AcREP21-32
AcREP21-3
AcREP21-36
AcREP21-10
AcREP21-39
AcREP21-46
AcREP21-14
AcREP21-15
AcREP21-31
AcREP21-16
AcREP21-17
AcREP21-38
AcREP21-45
AcREP21-47
AcREP21-9
AcREP21-43
AcREP21-7
AcREP21-21
AcREP21-12
AcREP21-25
AcREP21-34
AcREP21-5
AcREP21-19
AcREP21-20
AcREP21-28
AcREP21-33
AcREP21-29
AcREP21-30
AcREP21-48
AcREP21-13
AcREP21-22
AcREP21-23
AcREP21-37
AcREP21-8
AcREP21-27

```
AtREP21-2
AtREP21-11
AtREP21-40
AtREP21-4
AtREP21-24
AtREP21-41
AtREP21-1
AtREP21-42
AtREP21-26
AtREP21-6
AtREP21-35
AtREP21-18
AtREP21-44
AtREP21-32
AtREP21-3
AtREP21-36
AtREP21-10
AtREP21-39
AtREP21-46
AtREP21-14
AtREP21-15
AtREP21-31
AtREP21-16
AtREP21-17
AtREP21-38
AtREP21-45
AtREP21-47
AtREP21-9
AtREP21-43
AtREP21-7
AtREP21-21
AtREP21-12
AtREP21-25
AtREP21-34
AtREP21-5
AtREP21-19
AtREP21-20
AtREP21-28
AtREP21-33
AtREP21-29
AtREP21-30
AtREP21-48
AtREP21-13
AtREP21-22
AtREP21-23
AtREP21-37
AtREP21-8
AtREP21-27
```

AtREP21-2 ----------AGAAATATTTT----ATAA
AtREP21-11 ----------AAATGTATTTT----ATAA
AtREP21-40 ----------AGATTTTTTTTTT-TATAA
AtREP21-4 ----------AGAAATATTTT----TATAA
AtREP21-24 ----------AAATTATATGTT----ACTA
AtREP21-41 ----------AAATTATATGTT----ACTA
AtREP21-1 ----------TATATT----TATAA
AtREP21-42 ----------TATATT----TATAA
AtREP21-26 ----------TATATT----TATAA
AtREP21-6 ----------TATATT----TATAA
AtREP21-35 ----------TATATT----TATAA
AtREP21-18 ----------TATATT----TATAA
AtREP21-44 ----------TATATT----TATAA
AtREP21-32 ----------TATATT----TATAA
AtREP21-3 ----------TATATT----TATAA
AtREP21-36 ----------TATATT----TATAA
AtREP21-10 ----------TATATT----TATAA
AtREP21-39 ----------TATATT----TATAT
AtREP21-46 ----------TATATT----TATAA
AtREP21-14 ----------TATATT----TCTAA
AtREP21-15 ----------TATATT----TATAA
AtREP21-31 ----------TATATT----TATAA
AtREP21-16 ----------TATATT----TATAA
AtREP21-17 ----------TTTATT----TATAA
AtREP21-38 ----------TATATT----TATAA
AtREP21-45 ----------TATATT----TATAA
AtREP21-47 ----------TATATT----TATAT
AtREP21-9 ----------TATATT----TATAA
AtREP21-43 CTTATAATCATAAAAAATATTACACCAATTTTTGTGAATCTTACATTTTTGTAAATACTTTACTTTCACTAAAATTACAGACACATACTAATTAATACACATTCTTCTTAACACTAAATAACAACCATAAAACAAAAATCCCAATAATTTTAGTTATGTTTGATTCAAAAAAATATATATT----TATAA
AtREP21-7 ----------TATATT----TATAA
AtREP21-21 ----------TATATT----TATAA
AtREP21-12 ----------TATATT----TATAA
AtREP21-25 ----------TATATT----TATAA
AtREP21-34 ----------TATATT----TATAA
AtREP21-5 ----------TATATT----TATAA
AtREP21-19 ----------TATATT----TATAA
AtREP21-20 ----------TATATT----ATAA
AtREP21-28 ----------TATATT----ATTT
AtREP21-33 ----------TTTATT----TATAA
AtREP21-29 ----------TATATT----TATAA
AtREP21-30 ----------TATATT----TATAA
AtREP21-48 ----------TATATT----TATAA
AtREP21-13 ----------TATATT----TATAA
AtREP21-22 ----------TTTATT----TATAA
AtREP21-23 ----------TGTATT----TATAA
AtREP21-37 ----------TATATT----TATAA
AtREP21-8 ----------GGACATATCTTAAATTTTAT
AtREP21-27 ----------TGACACATCTTAAAATTTAT

```
AtREP21-2   ATTTACAATATAAAAA----------TATAATCAAAATACTTAATTAA----------AAAAC-----ATATAAAACTGAGGCCCCCGGTAAACCGCGGGTTAAATCCTAG
AtREP21-11  ATTTACAATATAAAA----------TATAACCAAA-TACTTAATTAA----------AAACCC----ATATAAAACCGAGGCCCGCTGTAAATCGCGGGTTAAAACCTAG
AtREP21-40  ATTTACAGTATAAAA----------TATAATCAAA-TACTTAATTAA----------AA--CT----ATATAAAACCGAGACCCGCGGTAAACCGCGGGTTAAACCCTTG
AtREP21-4   ATTTATAATATAAAA----------TATAATCAAA-TACTTAATTAA----------AAA-CT----ATATAAAACCGAGGTCCGCGGTAAACCGCGAGTTAAACCCTAG
AtREP21-24  TTTTTTAATATATACA----------CACACTAATTTATTTTAATAAA----------AAATCT----ATA-AAAGCTTTCGCCCGCGGGAAACCGCGGGTTTAATCCTAG
AtREP21-41  TTTTTTAATATATACA----------CACACTAATTTATTTAATAAA----------AAATCT----ATA-AAAGCCTTCGCCCGCGGGAAACCGCGGGTTTAATCCTAG
AtREP21-1   GTTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-42  GTTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-26  GTTTAGGATATCAAA----------TTTATTTCCAA-TATTTTGTATA----------AAAACC----AAATAAACCGATTGTTCACGGTAAACCGCGGGTTAAAACCTAG
AtREP21-6   GTTTAAAATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AAAACC----AAATAAACCGATCGTCTGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-35  GTTTAAA-TATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACA----AAATAAACCGATGGCCCGTATAAACCGCGGGTTAAAACCTAG
AtREP21-18  GTTTAAAATATCAAA----------TTTCTTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGTCTGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-44  GTTTAAGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAATTGATTGTCCCGCGGTAAACCGCGAGTTAAAACCTAG
AtREP21-32  GTTTATGATATCAAA----------CTTATTTCAAA-TATTTAATATA----------AGAACC----AAATAAATTGATTGTCCGCGGTAAACCGCGGGTTAAAACCTTA
AtREP21-3   GTTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AAAACC----AAATAAACCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-36  GTTTAGGATATCAAA----------TTTAATTCAAA-TATTTAATATA----------AGAACT----AAATAAACCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-10  ATTTAAGATATTTAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGTCCCGCGGTAAACTGCTGGTTAAAGCCTAG
AtREP21-39  GTTTAAAATATCATA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGTCCGCGGTAATCCGCGGTTAAATCCTAG
AtREP21-46  ATTTAAAATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AAAACC----AAATCTACCGATTGCCCGCGGTTGCCGCGGGTTAAATTCTAG
AtREP21-14  GTTTAAAATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACT----AAATAAACCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-15  GTTTAAAATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AAAATC----AAATAAACCGATTGTCCGCGGTTTACCGCGGGTTAAAACCTAG
AtREP21-31  ATTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AAAACC----AAATAAACCGATTGTCCGCACGTAAATCGCGGGTTAAAACCTAG
AtREP21-16  GTTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGTCCGCGGTAAACCGCGTGTTAAAACCTAG
AtREP21-17  GTTTAGGATATCAAA----------GTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGCCCGCGGTAAACCGCGGTAGTTAAAACCTAG
AtREP21-38  GTTTAAGATATCAAA----------TTTATTTCATA-TATTTTAGTATA----------AGAACC----AAATAAACCGATTGCCCGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-45  GTTTAAGATATCAAA----------TTTATTTCAAA-TATTTTAGTATA----------AGAACC----AAATAAACCGATTGCCCGCAGGTTAAAACCTA
AtREP21-47  GTTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGCCCGCGGTAAACCGCGGGTTAAAACCTAG
AtREP21-9   GTTTAAGATATCAAG----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGTCCGCGGTAAACCACAGGTTAAATCCTAG
AtREP21-43  GTATAGGATATCAAA----------TTTAT--AAAATATTTAGTATA----------AGAACC----AAATAATACGATTGCATGCGGTTAAACCGCGGGTTAAATCCTAG
AtREP21-7   GTTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGCCCGCGGTAAATTGCGCGGGTTAAAACCTAG
AtREP21-21  GTTTAGAATATCAAA----------TTTATTTCGAA-TATTTAGTATA----------AAACCC----AAATAAACCGATTGCCCGCTGTAAACCGCGGGTTAAAACCTAG
AtREP21-12  ATTTAAGATATCAAA----------TTCATTTA-AATATTTAGTATA----------AGAACC----AAATAAACCGATTGCCCGCGGTAAACCGCGGGTTAAATCCTAG
AtREP21-25  ATTTAAGATATCAAA----------TTCATTCA-AATATTTAGTATA----------AAAACC----AAATAAACCGATTGCCCGCGGTAAACCGCGGGTTAAATCCTAG
AtREP21-34  ATTTAAAATATCAAA----------TTTATTCA-AATATTTAGTATA----------AAAACC----AAATAAACCGATTGTCCGCGGTTTACCGCGGATTAAATCCTAG
AtREP21-5   GTTTAGGATATCAAA----------TTCATTCA-AATATTTAGTATA----------AGAATC----ACATAAACCGATTGCCCGCGGTAAACCGCGGGTTAAATCCTAG
AtREP21-19  GTTTAGGATATCAAA----------TTCATTCA-AATATTTAGTATA----------AGAATC----AAATAAACCGATTGCCCGCAGTAAACCGCGGGTTAAATCTTAG
AtREP21-20  GTTTAGGATATCAAA----------TTTATTCAAAATATTTAATATA----------AGAATT----AAATAAACCGATTGTCGTGGTAAACCGCGGGTTAAATCCTAG
AtREP21-28  ATTTC----------------------AATATTTAGTATA----------AGAACT----AAATAAACCGATTGCCCGCGGTAAACCGCGGGTTAAATCCTAG
AtREP21-33  GTTTAAGATATCAAA----------TTTATTCA-AATATTTAGTATA----------AAAACC----AAATAAACCGATTGCCTGCGGTAAACCGCGGGTTAAATCCTAG
AtREP21-29  GTTTAGGATATCAAA----------TTTATTCA-AATATTTAGTATA----------AGAACC----AAATAAACCGATTGCCCGCGGTAAACCGCGGGTTAAATCCTAG
AtREP21-30  GTTTAGGATATCAAA----------TTTACTCA-AATATTTAGTATA----------AAAACC----AAATAAACCGATTGCCCGCGGTAAACCGTGGGTTAAATCTTAG
AtREP21-48  GTTTAGGATATCAAA----------TTTACTCA-AATATTTAGTATA----------AAACCT----AAATAAACCGATTGCCCGCGGTAAACCGTAGTTAAATCCTAG
AtREP21-13  GTTTAGGATATCAAA----------TTTACTCA-AATATTTAGTATA----------AAAACC----AAATAAACCGATTGCCCGCGGTAAACCGTGAGTTAAATCCTTG
AtREP21-22  GTTTAAGATATCAAA----------TTTATTCA-AATATTTAGTATA----------AAAACC----AAATAAACCGATTGTCCGCGGTAAACCGCGGGTTAAATCCTAG
AtREP21-23  GTTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATA----------AGAACC----AAATAAACCGATTGCCGCGGTAAACCGCGGCTTAAGTCCTAG
AtREP21-37  GTTTAGGATATCAAA----------TTTATTTCAAA-TATTTAGTATATATA----------AGAACC----AAATAAACTGATTGCCCGTGGTAAACCGCGATGTTAAATCCTAG
AtREP21-8   AATTACAATACTAATTTGGACGTAACACGAGTTAAACTAAATTATATAATTCTAAATACTATTTTGACAACTGTTTAAATAAAAAA--TCTCCCGCGGTAAACCGCGAGTTTAATCCTAG
AtREP21-27  ATTACAATACTAAATTGGACGCAACACGGGTTAAACTAAATTATATAATTCAAATACTATTTTGACAACTGTTTAAATAAAAAA--TCTCCCGCGGTAAACCGCGGGTTTAATCCTAG
```

5.3 Annexe 3 : Séquences consensus des domaines contenus dans AtREP21

Les consensus présentés ci-dessous au format FASTA sont issus des matrices HMM de chaque domaine. Les lettres présentées en majuscules sont des nucléotides significativement conservés (au sens statistique du terme) à leur position.

```
>Domaine_1
aggtaactaggatttaacccgcggtttaccgcgggcaatcggtttatttagttcttatactaaatattgaaataaataatataaat
tttttaatcaaacataactaaataataattaggatttttttattgttatttaatgataagaggaatgtgtttaaatttttttggt
tgttatttacttatttttgttgaacagtttagattgaagagaattaattttattgaacggtttagcttgaatcggatcatataaag
tttaagagattgttaagaccaaattacctttaataatttttgaatagcaaagaagaaaatccaaaattgaaatcaaaattgaggcca
aattagacaaattttcaatttgtacctccttttttaataataaga

>Domaine_2
aggttaatttggtcaactcaatttattaaagcttataggatccgcgtttt
ta

>Domaine_3
tattccttaaacacactatatttcaaatttaaataa

>Domaine_4
tacataataggttaccatcaagta

>Domaine_5
caaaatagacaatcttaaccg

>Domaine_6
aatcaatctaaaataaaccaaaaata

>Domaine_7
atgcaatcttaaccataacacgatta

>Domaine_8
ttcaacaatattggtatctctacttcattttggtat

>Domaine_9
ataaggaatctatgtcatttTtaaaatcaaaatAtaatcAcattccaTcaataaAaacagtatAttcAAtcacgaaatcAAtAAtc
aAtAAtAtctaaAtcAtaTaTTacgattcttcacTaATCAcTAgAa

>Domaine_10
aatcgacatagacgagcttattcgat

>Domaine_11
tctttattttgccaaaaaactgtataacctgtcaaatagttaaaaaaaacacttataatcataaacacttatta

>Domaine_12
ctcaaaaagaattataatcacaaaa

>Domaine_13
tctaaaactaaccataatcaactaaa

>Domaine_14
caatatttaccacttatctaaatata
```

>Domaine_15
gccaatcattatgaactaaaccaaattcgaatcattagtgaggatccaatctttgatttgcaacttatcctgcaaaatattttcga
a

>Domaine_16
atgaaaaatattatgttggcaaacaatattaagagctcgcg

>Domaine_17
accctagcctccttcatcactataaatatcaaagtcattctcatcacaaaccaaattgagatcaagaacgattcgatggttatatc
ataatcaattcatgatagtgaatgattgtttgcataggtattcatgacgaagtggatgttcttctttgatgttgatgagtttttac
atgttccggtgaaagagacgatttcgtctgtgatggaatctttggaggaatattttcagtttactattgaactgatgcctatgagt
agtcgggtttgttattccggtgatggcccggccagaacttacaggtaacattcataactttgttgattttaatgattgaacgttaa
ccaaagcccatgtgttgtttat

>Domaine_18
tgagtgacttagaatattagtcacaa

>Domaine_19
aaatatttcaaaaaacacttataataacaaaaaga

>Domaine_20
gttttttgaaggtgacatataatatcaataagcatctgattctgtcaacacttggcctgtaaaagtctttgtgcattctgtattag
tctcctatgtcggtcacgatctttgcttttattgaacatcacatgatattagtgagaaatgttgtaaagattatttcttttgtttg
taccaaatattgaattgtaattggtttgtatattttctgtttttaggaaggaacatatcggttattgatttggtttagttcaggga
ctaatctaaatggtatttggttttgtgtttcagaaaatggggaattgagaaactggcatatagagacgtcaagaaggttccaaga
cgggatcggaaatacgcggtccagccggagaatgtattcgcgataggcgtacacatgtctcagaatctacaagggaagacatagca
caaggctgagagcaaaatccgttacttccattaccatggttcgatctctcagcgccgcgagcacgcactccttatgtgctagacac
tacaatccgtgatgttggccttgctgtgagaacgttcgagttgagaacgattggtgatcggctgctacggacaagacaatgaagag
atgaaaaacatgaatagtgaatctaatcaatctttagaaagaagaattaggtgttaagatgatttactttgcattgttttcttttg
ggggtatattcttttattgtatcatataatttgggtaatgggatcattaatacagcttcaaaatactctttagtatatatattctg
tatgatttagaaaaaaggtctcctcataaatgttattgtctttgttttgccaaaaataaaagataaaacataccaactttggtttt
taaa

>Domaine_21
gaacttataatcaaaaaaagaattac

>Domaine_22
aataattcataattatgcatatggaa

>Domaine_23
aagtgaaataatgttatttttttatcttcactttaaa

>Domaine_24
gacatataattataaataacataattaatatggtcttt

>Domaine_25
accaxttttaataatattacatttttgttactacttt

>Domaine_26
acttatatttaaattacataaacata

>Domaine_27
taacaatcttaacggtccattaaaactttccaaa

>Domaine_28
ttccatcccaaaaagaattataatc

>Domaine_29
tttaaacacattaaaacacattattc

>Domaine_30
caatctaaattattatttgatttaat

>Domaine_31
tcaaaatgatttccaaatcaatcatt

>Domaine_32
aacaaataccgacctgcaaaaacggg

>Domaine_33
taatcagtaaatagactccatactaataaaaga

>Domaine_34
tcatatattacgattcttactaatc

>Domaine_35
caaaaaaaaatcccaattattatttagttatgtttg

>Domaine_36
taagcatattcgacaaacaaaagcatattccacaaatggtttacaaaatcgacaaatcgagtcttcatcaccaactaaacttattc
tcttcaatctaatctattcataactatctacacaatattcagtattaatctaaatttaaccaatcattaaaatttaaaccattcct
aaagattttatgctttctttacaagattgttacaatattcatatggtctttttttgccaaaaaagtgtataacctgtcaaatagtt
taaaaaaaaacacttataatcataaaaaatattacaccaattttgtgaatcttacatttttgtaaatactttacttttcactaaaat
tacagacacataactaattaatacacatttttttaacacattcttcttaacactaaataacaaccaaaacaaaaatcccaataatta
tttagttatgtttgattcaaaaaaa

>Domaine_37
aacacacactaatttatttaataaa

>Domaine_38
taaattggacgcaacacgagttaaactaaattatataattcaaatactattttg

>Domaine_39
aaatctataaaagccttcgcc

>Domaine_40
acaactgtacaaataaaaaatctcc

>Domaine_41
aaaTAtTtagTATAAgAAccAAATaaaccGATtGcc

>Domaine_42
CGcggTaaacCGcgggTTaAAtccta

>Domaine_43
cgcggtaaactgctggttaaagttat

>Domaine_44
TccctTTATTAtTAaAAgggAAgTAcAAAtTgAAattTtgtctAATtTggccTcAATTTtgATTTcAAtTTTggATTTTgTtcttt
gctAttcAaaaTc

>Domaine_45

```
aaaatcaatattcgatatatcaaaa

>Domaine_47
caatctcctaaactttatatgatccgat

>Domaine_48
atagttattaattttacatattttt

>Domaine_49
gttagaatttgtatttatttaatagag

>Domaine_50
ttttttcaaagatttaaaaaaactaaatata

>Domaine_51
ttctaatcactaaataacaaccaatcaaaaa

>Domaine_52
tatttaaAaatttataTtTAtAAgTTTAxaaTAtcAaAtt

>Domaine_55
aatcccaattattatttagttatgtttgatta

>Domaine_56
aaatctcaatcattatttataaaaatttaaaa

>Domaine_57
attcccTTcAATcTAAaccgtTcAAcAAAAatA

>Domaine_58
ttatattatttatttcaatatttagtataa

>Domaine_59
tttatatgATccgaTTcAATctAAACcgTTcAAtAAaAtTA

>Domaine_46-ci
tgatggaaaatttcttggtcaaaattaaa

>Domaine_47-ci
aaccatataaaaccgaggcc

>Domaine_51-ci
atttcgggtgggttaggagtaat

>Domaine_52-ci
taaattacatacacattaxtaataatacacata

>Domaine_54-ci
aatgagtaaaatggaattgttaatga

>Domaine_58-ci
cttcattttgatatctaaatatgagcata

>Domaine_59-ci
aatattggagcggattagaatttaca
```

```
>Domaine_49-i
atagacgggtgagttatgtttaat

>Domaine_51-i
actagaaaatccacaaacttattcac

>Domaine_53-i
gtaaaatcacacattttccggatattatttaaac

>Domaine_55-i
AaAATcattaAgGgTAATTTGGTctTaacaa

>Domaine_56-i
gggcttttgatcacatgtaaaac

>Domaine_58-i
aaaatataaccaaatacttaattaa

>Domaine_60-i
ccttggacacatctaaaattataattacaatac
```

5.4 Annexe 4 : Tableau d'informations des AtREP3 insérés dans les promoteurs de gènes

Nous avons limité notre tableau au AtREP3 présent à moins de 1kb du codon START des gènes à leur proximité. Le gène de l'ADC1 est le locus At2G16500.

N°	Début	Fin	Sens	Distance	Locus	Début	Fin	Sens
17	499373	501371	-	857	At1G02450	497976	498516	-
1	3413554	3414840	+	710	At1G10390	3407025	3412844	-
3	7035181	7037279	+	317	At1G20320	7033647	7034864	-
18	8489403	8490182	-	441	At1G23990	8490623	8492719	+
4	9816197	9818260	+	669	At1G28120	9812952	9815528	-
19	9865331	9867400	-	738	At1G28230	9862060	9864593	-
6	12364115	12365971	+	896	At1G34020	12366867	12369158	+
23	13015829	13017601	-	666	At1G35400	13018267	13019029	+
7	13058801	13060873	+	743	At1G35490	13061616	13063424	+
25	16999931	17001394	-	595	At1G44960	17001989	17004169	+
11	18793771	18794515	+	459	At1G50720	18792848	18793312	-
26	19330906	19332928	-	355	At1G51980	19327071	19330551	-
27	19468759	19470617	-	745	At1G52270	19467354	19468014	-
28	19771658	19773500	-	642	At1G53040	19767978	19771016	-
129	19997348	19999401	-	450	At1G53580	19994943	19996898	-
129	19997348	19999401	-	717	At1G53590	20000118	20003933	+
14	22453596	22455669	+	472	At1G60970	22451513	22453124	-
14	22453596	22455669	+	569	At1G60980	22456238	22457805	+
131	22625444	22625969	-	671	At1G61330	22626640	22628192	+
29	22995597	22997725	-	253	At1G62230	22994640	22995344	-
29	22995597	22997725	-	523	At1G62240	22998248	22999136	+
16	23879809	23880930	+	858	At1G64340	23881788	23882718	+
32	26173058	26175142	-	660	At1G69580	26175802	26177275	+
33	30258034	30260414	-	838	At1G80470	30261252	30262808	+
37	542307	544163	+	543	At2G02140	544706	545424	+
132	707441	710071	+	640	At2G02610	710711	712818	+
44	2962291	2962801	-	149	At2G07150	2962950	2964179	+
45	3197360	3213350	-	659	At2G07635	3214009	3217560	+
40	7118866	7120009	+	718	At2G16410	7120727	7121341	+
134	7160877	7161817	-	456	At2G16500	7157704	7160421	-
48	7820484	7821772	-	273	At2G17960	7819462	7820211	-
49	11292193	11294238	-	886	At2G26530	11289853	11291307	-
136	12100559	12101501	-	322	At2G28320	12101823	12106721	+
42	13273992	13276081	+	355	At2G31130	13270758	13273637	-
42	13273992	13276081	+	922	At2G31140	13277003	13278872	+
51	13913744	13915797	-	341	At2G32790	13912567	13913403	-

N°	Début	Fin	Sens	Distance	Locus	Début	Fin	Sens
53	1571386	1572131	+	593	At3G05440	1572724	1573272	+
54	3716348	3717217	+	438	At3G11750	3715077	3715910	-
56	10321649	10322677	+	320	At3G27831	10320718	10321329	-
59	15948273	15950357	+	814	At3G44240	15951171	15951890	+
60	16430129	16431438	+	592	At3G44950	16432030	16432271	+
73	17072371	17074213	-	197	At3G46382	17074410	17075090	+
75	17286728	17288798	-	823	At3G46910	17289621	17290637	+
63	17740925	17742216	+	249	At3G48040	17742465	17744477	+
64	18029766	18030911	+	130	At3G48630	18029268	18029636	-
65	18423139	18424283	+	392	At3G49660	18424675	18426379	+
66	18595721	18596085	+	203	At3G50110	18591592	18595518	-
67	20797378	20799425	+	245	At3G56010	20799670	20800903	+
67	20797378	20799425	+	605	At3G56000	20794308	20796773	-
68	22129590	22131633	+	372	At3G59870	22127138	22129218	-
68	22129590	22131633	+	431	At3G59880	22132064	22133127	+
83	370788	372085	-	932	At4G00885	369856	369991	-
77	683213	684357	+	630	At4G01575	681868	682583	-
92	16863268	16865074	-	419	At4G35520	16865493	16871769	+
92	16863268	16865074	-	891	At4G35510	16860457	16862377	-
93	4312439	4314868	+	424	At5G13450	4310317	4312015	-
106	4780213	4782230	-	719	At5G14790	4782949	4785808	+
107	5427959	5430064	-	484	At5G16580	5425892	5427475	-
108	7153186	7154440	-	320	At5G21060	7148980	7152866	-
98	11025340	11027899	+	389	At5G29000	11022087	11024951	-
116	15621899	15623830	-	171	At5G38980	15624001	15624579	+
99	16373703	16375771	+	186	At5G40830	16371608	16373517	-
100	17382253	17384350	+	766	At5G43280	17385116	17386542	+
118	19757741	19759842	-	725	At5G48670	19756051	19757016	-
141	20298271	20299209	-	182	At5G49880	20299391	20305310	+
119	20697542	20699631	-	874	At5G50830	20700505	20702081	+
120	21793013	21794257	-	76	At5G53600	21794333	21794668	+
121	22248178	22249489	-	452	At5G54710	22244800	22247726	-
122	24492517	24493816	-	800	At5G60840	24490560	24491717	-
103	25281737	25282562	+	690	At5G62950	25279458	25281047	-
104	25921958	25922840	+	469	At5G64800	25923309	25923629	+

Bibliographie

[1] M.I. Abouelhoda, S. Kurtz, and E. Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, 2 :54–86, 2004.

[2] S. Aerts, P.V. Loo, G. Thijs, H. Mayer, Martin. de, . Moreau, and Moor. De. Toucan 2 : the all-inclusive open source workbench for regulatory sequence analysis. *Nucl. Acids Res.*, 33 :W393–W396, 2005.

[3] SF. Altschul, TL. Madden, AA. Schäffer, J. Zhang, Z. Zhang, W. Miller, and DJ. Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res.*, 25 :3389–3402, 1997.

[4] J.O. Andersson. Lateral gene transfer in eukaryotes. *CMLS and Cell. Mol. Life Sci.*, 62 :1182–1197, 2005.

[5] R. Andonov, D. Lavenier, P. Veber, and N. Yanev. Dynamic programming for lr-pcr segmentation of bacterium genomes. *Concurrency and Computation : Practice and Experience*, 17 :1657–1668, 2005.

[6] C. Andre, P. Vincens, JF. Boisvieux, and S. Hazout. Mosaic : segmenting multiple aligned dna sequences. *Bioinformatics.*, 17 :196–197, 2001.

[7] C. Augé-Gouillou, H. Notareschi-Leroy, P. Abad, G. Periquet, and Y. Bigot. Phylogenetic analysis of the functional domain of mariner-like element (mle) transposases. *Mol. Gen.Genet.*, 264 :506–513, 2000.

[8] F.M. Ausubel. Summaries of national science foundation-sponsored arabidopsis 2010 projects and national science foundation-sponsored plant genome projects that are generating arabidopsis resources for the community. *Plant Physiology.*, 129 :394–437, 2002.

[9] RK. Azad, JS. Rao, W. Li, and R. Ramaswamy. Simplifying the mosaic description of dna sequences. *Phys Rev E Stat Nonlin Soft Matter Phys.*, 66 :031913, 2002.

[10] K.M. Konwar B. DasGupta and, I.I. Mandoiu, and A.A. Shvartsman. Dna-bar : distinguisher selection for dna barcoding. *Bioinformatics*, 21 :3424–3426, 2005.

[11] TL. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with meme. *Proc Int Conf Intell Syst Mol Biol*, 3 :21–29, 1995.

[12] Z. Bao and SR. Eddy. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.*, 12 :1269–1276, 2002.

[13] J.A. Bedell, I. Korf, and Gish. and. Maskeraid : a performance enhancement to repeatmasker. *Bioinformatics*, 16 :1040–1041, 2000.

[14] G. Bejerano, CB. Lowe, N. Ahituv, B. King, A. Siepel, SR. Salama, EM. Rubin, WJ. Kent, and D. Haussler. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441 :87–90, 2006.

[15] J.L. Bennetzen. Transposable elements and gene creation and genome rearrangement in flowering plants. *Curr. Op. Genet. Dev.*, 15 :621–627, 2005.

[16] I. Bernales, M.V. Mendiola, and F. de la Cruz. Intramolecular transposition of insertion sequence is91 results in second-site simple insertions. *Mol. Microbio.*, 33 :223–234, 1999.

[17] P. Bernaola-Galvan, R. Roman-Roldan, and JL. Oliver. Compositional segmentation and long-range fractal correlations in dna sequences. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics.*, 53 :5181–5189, 1996.

[18] F. Bourgeois and J-C. Lasalle. An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*, 14 :802–806, 1971.

[19] J. Brandt, S. Schrauth, AM. Veith, A. Froschauer, T. Haneke, C. Schultheis, M. Gessler, Leimeister., and JN. Volff. Transposable elements as a source of genetic innovation : expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene*, 345 :101–11, 2005.

[20] R. Britten. Transposable elements have contributed to thousands of human proteins. *Proc Natl Acad Sci U S A*, 103 :1798–803, 2005.

[21] T.A. Brown. Genomes 3rd edition. *ISBN 1-85996-228-9*, 2006.

[22] A. Bruce, J. Alexander, L. Julian, R. Martin, R. Keith, and W. Peter. Molecular biology of the cell 4th ed. *ISBN 0-8153-3218-1*, 2002.

[23] S. Brunner and G. Pea. Origins and genetic organization and transcription of a family of non-autonomous helitron elements in maize. *The Plant Journal*, 43 :799–810, 2005.

[24] M.B. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biol.*, 5 :E201, 2003.

[25] P. Bundock and P. Hooykaas. An arabidopsis hat-like transposase is essential for plant development. *Nature*, 436 :282–284, 2005.

[26] C. Burge and S. Karlin S. Prediction of complete gene structures in human genomic dna. *J. Mol. Biol.*, 268 :78–94, 1997.

[27] P. Capy, C. Bazin, and D. Higuet. Dynamics and evolution of transposable elements. *Springer and Landes Biosciences and Library of Congress and Austin and Texas*, 1998.

[28] CM. Carlson, JL. Frandsen, N. Kirchhof, RS. McIvor, and DA. Largaespada. Somatic integration of an oncogene-harboring sleeping beauty transposon models liver tumor development in the mouse. *P.N.A.S.*, 22 :17059–64, 2005.

[29] E. Casacuberta and ML. Pardue. Transposon telomeres are widely distributed in the drosophila genus : Tart elements in the virilis group. *P.N.A.S.*, 100 :3363–3368, 2003.

[30] F. Casals, M. Cáeres, M.H. Manfrin, J. González, and A. Ruiz. Molecular characterization and chromosomal distribution of galileo and kepler and newton and three foldback transposable elements of the drosophila buzzatii species complex. *Genetics*, 169 :2047–2059, 2005.

[31] A. Caspi and L. Pachter. Identification of transposable elements using multiple alignments of related genomes. *Genome Res.*, 16 :260–270, 2006.

[32] J.D. Choi, A. Hoshino, K. Park, I. Park, and S. Iida. Spontaneous mutations caused by a helitron transposon, hel-it1, in morning glory, ipomoea tricolor. *The Plant Journal*, 49 :924–934, 2007.

[33] N. Chomsky. Syntactic structures. *Mouton and Gravenhage and Netherlands*, 1957.

[34] J. Collado-Vides. Grammatical model of the regulation of gene expression. *P.N.A.S.*, 89 :9405–9409, 1992.

[35] C. Conte, B. Dastugue, and C. Vaury. Promoter competition as a mechanism of transcriptional interference mediated by retrotransposons. *The EMBO Journal*, 21 :3908–3916, 2002.

[36] G.E. Cooper and R.E. Hausman. The cell : A molecular approach and 4th edition. *ISBN 0-8153-3218-1*, 2006.

[37] R. Cordaux, DJ. Hedges, SW. Herke, and MA. Batzer. Estimating the retrotransposition rate of human alu elements. *Gene*, 24 :134–137, 2006.

[38] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. The set-covering problem, introduction to algorithms, second edition. *MIT Press and McGraw-Hill*, ISBN 0-262-03293-7. Section 35.3 :1033–1038, 2001.

[39] N.L. Craig, R. Gragie, M. Gellert, and A.M. Lambowitz. Mobile dna ii second edition. *ASM Press*, 2002.

[40] M. Crochemore and C. Hancart. Algorithmique du texte. *ISBN 2-7117-8628-5.*, 2001.

[41] M. Crochemore and R. Vérin. On compact suffix dawg. *In J. Mycielski and G. Rozenberg and A. Salomaa and editors and Mathematical Logic and Theoretical Computer Science and Lecture Notes in Computer Science. Springer-Verlag*, 1997.

[42] A. Cultrone, Y.R. Dominguez, C. Drevet, C. Scazzocchio, and R. Fernandez-Martin. The tightly regulated promoter of the xana gene of aspergillus nidulans is included in a helitron. *Mol Microbiol.*, 63 :1577–1587, 2007.

[43] A.C.E. Darling, B. Mau, F.R. Blattner, and N.T. Perna. Mauve : Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, 14 :1394–1403, 2004.

[44] SM. Daskalova, NW. Scott, and MC. Elliott. Folbos and a new foldback element in rice. *Genes Genet Syst*, 80 :141–145, 2005.

[45] R.V. Davuluri, H. Sun, S.K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz, and E. Grotewold. Agris : Arabidopsis gene regulatory information server and an information resource of arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, 4 :E25, 2003.

[46] V de Lagemaat, JR. Landry, DL. Mager, and P. Medstrand. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.*, 19 :530–536, 2003.

[47] Pilar. del, I. Bernales, MV. Mendiola, and la. de. Single-stranded dna intermediates in is91 rolling-circle transposition. *Molecular Microbiology*, 39 :494–501, 2001.

[48] A.L. Delcher, S. Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg. Alignment of whole genomes. *Nucleic Acids Research.*, 27 :2369–2376, 1999.

[49] J.M. Deragon. La relation sine/line : un exemple de parasitisme moléculaire ?. *Médecine/sciences* ., 17 :103–106, 2001.

[50] M. Dewannieux and T. Heidmann. L1-mediated retrotransposition of murine b1 and b2 sines recapitulated in cultured cells. *J Mol Biol.*, 349 :241–247, 2005.

[51] X. Diao, M. Freeling, and D. Lisch. Horizontal transfer of a plant transposon. *PLoS Biology.*, 4 :119–128, 2006.

[52] E. Diaz-Aroca, M.V. Mendiola, J.C. Zabala, and De. and. Transposition of is91 does not generate a target duplication. *J. Bact*, 169 :442–443, 1987.

[53] CR. Dietrich, F. Cui, ML. Packila, J. Li, DA. Ashlock, BJ. Nikolau, and PS. Schnable. Maize mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. *Genetics*, 160 :697–716, 2002.

[54] J. Drouaud, C. Camilleri, P.Y. Bourguignon, A. Canaguier, A. Berard, D. Vezon, S. Giancola, D. Brunel, V. Colot, B. Prum, H. Quesneville, and C. Mezard. Variation in crossing-over rates across chromosome 4 of arabidopsis thaliana reveals the presence of meiotic recombination "hot spots". *Genome Res.*, 2006 :106–114, 16.

[55] R. Druker, TJ. Bruxner, NJ. Lehrbach, and E. Whitelaw. Complex patterns of transcription at the insertion site of a retrotransposon in the mouse. *Nucleic Acids Res.*, 32 :5800–5808, 2004.

[56] CA. Dunn and DL. Mager. Transcription of the human and rodent spam1 / ph-20 genes initiates within an ancient endogenous retrovirus. *BMC Genomics.*, 6 :47–61, 2005.

[57] P. Durand, F. Mahe, A.S. Valin, and J. Nicolas. Browsing repeats in genomes : Pygram and an application to non-coding region analysis. *BMC Bioinformatics*, 7 :477–494, 2006.

[58] N.A. Eckardt. A new twist on transposons : The maize genome harbors helitron insertion. *The Plant Cell.*, 15 :293–295, 2003.

[59] S.R. Eddy. Profile hidden markov models. *Bioinformatics.*, 14 :755–763, 1998.

[60] RC. Edgar and EW. Myers. Piler : identification and classification of genomic repeats. *Bioinformatics.*, Suppl 1 :i152–i158, 2005.

[61] R. Ekins and F.W. Chu. Microarrays : their origins and applications. *Trends in Biotechnology.*, 17 :217–218, 1999.

[62] A. ElAmrani, L. Marie, A. Ainouche, J. Nicolas, and I. Couee. Genome-wide distribution and potential regulatory functions of atate and a novel family of miniature inverted-repeat transposable elements in arabidopsis thaliana. *Mol Genet Genomics.*, 267 :459–71, 2002.

[63] C. elegans Consortium. Genome sequence of the nematode caenorhabditis elegans. a platform for investigating biology. *Science*, 282 :2012–2018, 1998.

[64] SJ. Emrich, S. Aluru, Y. Fu, TJ. Wen, M. Narayanan, L. Guo, DA. Ashlock, and PA. Schnable. A strategy for assembling the maize (zea mays l.) genome. *Bioinformatics*, 20 :140–147, 2004.

[65] MB. Evgen'ev and IR. Arkhipova. Penelope-like elements : a new class of retroelements : distribution and function and possible evolutionary significance. *Cytogenet Genome Res.*, 110 :510–521, 2005.

[66] J. Fajkus, E. Sykorova, and AR. Leitch. Telomeres in evolution and evolution of telomeres. *Chromosome Res.*, 13 :469–479, 2005.

[67] AV. Favorov, MS. Gelfand, AV. Gerasimova, DA. Ravcheev, AA. Mironov, and VJ. Makeev. A gibbs sampler for identification of symmetrically structured and spaced dna motifs with improved estimation of the signal length. *Bioinformatics*, 21 :2240–2245, 2005.

[68] C. Feschotte, N. Jiang, and S.R. Wessler. Plant transposable elements : where genetics meets genomics. *Nature Genet. Reviews*, 3 :329–341, 2003.

[69] C. Feschotte and C. Mouchès. Evidence that a family of miniature inverted-repeat transposable elements (mites) from the arabidopsis thaliana genome has arisen from a pogo-like dna transposon. *Mol. Biol. Evol.*, 17 :730–737, 2000.

[70] C. Feschotte and S.R. Wessler. Treasures in the attic : Rolling circle transposons discovered in eukaryotic genomes. *P.N.A.S.*, 98 :8923–8924, 2001.

[71] D.J. Finnegan. Transposable elements. *Curr. Opin. Genet. Dev.*, 2 :861–867, 1992.

[72] RD. Fleischmann, MD. Adams, O. White, RA. Clayton, EF. Kirkness, AR. Kerlavage, CJ. Bult, JF. Tomb, BA. Dougherty, JM. Merrick, and al. et. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269 :496–512, 1995.

[73] A. Gionis and H. Mannila. Finding recurrent sources in sequences. *In pro-ceedings of the 7 th International Conference on Research in Computational Mo-lecular Biology (RECOMB). Apr 10-13 and Berlin and Germany*, 2003.

[74] MA. Grandbastien, C. Audeon, E. Bonnivard, JM. Casacuberta, B. Chalhoub, AP. Costa, QH. Le, D. Melayah, M. Petit, C. Poncet, SM. Tam, Sluys. Van, and C. Mhiri. Stress activation and genomic impact of tnt1 retrotransposons in solanaceae. *Cytogenet Genome Res. Review*, 110 :229–41, 2005.

[75] S. Gupta, A. Gallavotti, G.A. Stryker, and R.J. Schmidt. A novel class of helitron-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol. Bio.*, 57 :115–127, 2005.

[76] D. Gusfield. Algorithms on strings and trees and sequences : Computer science and computational biology. *ISBN-13 : 9780521585194*, 1997.

[77] C. Gutierrez. Geminivirus dna replication. review. *Cell. Mol. Life Sci.*, 56 :313–329, 1999.

[78] N. Halaimia-Toumi, N. Casse, M.V. Demattei, S. Renault, E. Pradier, Y. Bigot, and M. Laulier. The gc-rich transposon bytmar1 from the deep-sea hydrothermal crab and bythograea thermydron and may encode three transposase isoforms from a single orf. *J. Mol. Evol.*, 59 :747–760, 2004.

[79] E.R. Havecker and X. Gao. The diversity of ltr retrotransposons. *Genome Biology*, 5 :E225, 2004.

[80] C. Helgesen and P.R. Sibbald. Palm - a pattern language for molecular biology. *ISMB*, 1 :172–180, 1993.

[81] C. Huang, F. Morcos, S.P. Kanaan, S. Wuchty, D.Z. Chen, and J.A. Izaguirre. Predicting protein-protein interactions from protein domains using a set cover approach. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 4 :78–87, 2007.

[82] I. Hummel, G. Gouesbet, A. El Amrani, A. Ainouche, and I. Couée. Characterization of the two arginine decarboxylase (polyamine biosynthesis) paralogues of the endemic subantarctic cruciferous species pringlea antiscorbutica and analysis of their differential expression during development and response to environmental stress. *Gene*, 342 :199–209, 2004.

[83] H. International. Initial sequencing and analysis of the human genome. *Nature*, 409 :860–921, 2001.

[84] Rice. International. The map-based sequence of the rice genome. *Nature*, 436 :793–800, 2005.

[85] T. Inukai. Role of transposable elements in the propagation of minisatellites in the rice genome. *Mol. Genet. Genomics*, 271 :220–227, 2004.

[86] T. Inukai and Y. Sano. Sequence rearrangement in the at-rich minisatellite of the novel rice transposable element basho. *Genome*, 45 :493–502, 2002.

[87] C.A. Janeway, P. Travers, M. Walport, and . Shlomchik. Immunobiology : The immune system in health and disease. *ISBN 0815341016*, 2004.

[88] JM. Jeter, W. Kohlmann, and SB. Gruber. Genetics of colorectal cancer. *Oncology (Williston Park)*, 20 :269–276, 2006.

[89] N. Jiang, Z. Bao, X. Zhang, SR. Eddy, and SR. Wessler. Pack-mule transposable elements mediate gene evolution in plants. *Nature*, 431 :569–573, 2004.

[90] I. Jonassen. Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.*, 13 :509–522, 1997.

[91] C. Josefsson, B. Dilkes, and L. Comai. Parent-dependent loss of gene silencing during interspecies hybridization. *Curr. Biol.*, 16 :1322–1328, 2006.

[92] N. Juretic, T.E. Bureau, and R.M. Bruskiewich. Transposable element annotation of the rice genome. *Bioinformatics*, 20 :155–160, 2004.

[93] J. Jurka. Repeats in genomic dna : mining and meaning. *Curr. Opin. Struct. Biol.*, 8 :333–337, 1998.

[94] J. Jurka, V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase update and a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110 :462–467, 2005.

[95] J. Jurka, P.. Klonowski, V. Dagman, and P. Pelton. Censor - a program for identification and elimination of repetitive elements from dna sequences. *Computers and Chemistry*, 20 :119–122, 1996.

[96] V.V. Kapitonov. Self-synthesizing dna transposons in eukaryotes. *P.N.A.S.*, 103 :4540–4545, 2006.

[97] V.V. Kapitonov and J. Jurka. Molecular paleontology of transposable elements from arabidopsis thaliana. *Genetica*, 107 :27–37, 1999.

[98] V.V. Kapitonov and J. Jurka. Rolling-circle transposons in eukaryotes. *P.N.A.S.*, 98 :8714–8719, 2001.

[99] V.V. Kapitonov and J. Jurka. Molecular paleontology of transposable elements in the drosophila melanogaster genome. *P.N.A.S.*, 100 :6569–6574, 2003.

[100] V.V. Kapitonov and J. Jurka. Rag1 core and v(d)j recombination signal sequences were derived from transib transposons. *PLoS Biol*, 3 :E181, 2005.

[101] M. Kato, K. Takashima, and T. Kakutani. Epigenetic control of cacta transposon mobility in arabidopsis thaliana. *Genetics*, 168 :961–969, 2004.

[102] F. Kempken and F. Windhofer. The hatfamily : a versatile transposon group common to plants and fungi and animals and man. *Chromosoma*, 110 :1–9, 2001.

[103] M.G. Kidwell. Intraspecific hybrid sterility. *In Genetics and biology of Drosophila. Ed. Ashburner M. and Carlson H.L. and Thompson J.N. Academic Press Inc. London*, 3C :125–153, 1983.

[104] M.G. Kidwell. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115 :49–63, 2002.

[105] M.G. Kidwell and D.R. Lisch. Transposable elements as sources of variation in animals and plants. *P.N.A.S.*, 94 :7704–11, 1997.

[106] M.G. Kidwell and D.R. Lisch. Transposable elements and host genome evolution. reviews. *TREE*, 15 :95–99, 2000.

[107] O. Kohany, A.J. Gentles, J. Hankus, and J. Jurka. Annotation, submission and screening of repetitive elements in repbase : Repbasesubmitter and censor. *BMC Bioinformatics*, 7 :E474, 2006.

[108] F.A. Kondrashov and A.S. Kondrashov. Role of selection in fixation of gene duplications. *J Theor Biol.*, 239 :141–151, 2006.

[109] DA. Kramerov and NS. Vassetzky. Short retroposons in eukaryotic genomes. *Int. Rev. Cytol.*, 247 :165–221, 2005.

[110] E. Kravchenko, E. Savitskaya, O. Kravchuk, A. Parshikov, P. Georgiev, and M. Savitsky. Pairing between gypsy insulators facilitates the enhancer action in trans throughout the drosophila genome. *Mol. Cell Biol.*, 25 :9283–9391, 2005.

[111] J. Kreahling and BR. Graveley. The origins and implications of aluternative splicing. *Trends Genet.*, 20 :1–4, 2004.

[112] M. Kuhlmann, BE. Borisova, M. Kaller, P. Larsson, D. Stach, J. Na, L. Eichinger, F. Lyko, V. Ambros, F. Soderbom, C. Hammann, and W. Nellen. Silencing of retrotransposons in dictyostelium by dna methylation and rnai. *Nucleic Acids Res.*, 33 :6405–17, 2005.

[113] HW. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistic Quarterly, Kuhn's original publication*, 2 :83–97, 1955.

[114] A. Kumar. Plant retrotransposons. *Annual Review of Genetics.*, 33 :479–532, 1999.

[115] S. Kurtz. Foundation of sequence analysis. *Algorithmica*, 2001b.

[116] S. Kurtz, J.V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. Reputer : The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, 29 :4633–4642, 2001.

[117] S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S.L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biology*, 5 :R12, 2004.

[118] NC. Kyrpides. Genomes online database (gold 1.0) : a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, 15 :773–774, 1999.

[119] J. Lai, Y. Li, and J. Messing. Gene movement by helitron transposons contributes to the haplotype variability of maize. *P.N.A.S.*, 102 :9068–9073, 2005.

[120] S.K. Lal and Hannah. and. Helitrons contribute to the lack of gene colinearity observed in modern maize inbreds. *P.N.A.S.*, 102 :9993–9994, 2005.

[121] S.K. Lal, M.J. Giroux, V. Brendel, C.E. Vallejos, and Hannah. et. The maize genome contains a helitron insertion. *The Plant Cell*, 15 :381–391, 2003.

[122] J. Laufs, W. Traut, F. Heyraud, V. Matzeit, S.G. Rogers, J. Schell, and Gronenborn. In vitro cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus. *P.N.A.S.*, 92 :3879–3883, 1995.

[123] Q.H. Le, S. Wright, Z. Yu, and T. Bureau. Transposon diversity in arabidopsis thaliana. *P.N.A.S.*, 97 :7376–7381, 2000.

[124] A. Lefebvre, T. Lecroq, H. Dauchel, and J. Alexandre. Forrepeats : detects repeats on entire chromosomes and between genomes. *Bioinformatics*, 19 :319–326, 2002.

[125] M.P. Lefranc, V. Giudicelli, Q. Kaas, E. Duprat, J. Jabado-Michaloud, D. Scaviner, C. Ginestoux, O. ClÃ©ment, D. Chaume1, and G. Lefranc. Imgt, the international immunogenetics information systemimgt, the international immunogenetics information system. *Nucleic Acids Research*, 33 :D593–D597, 2005.

[126] E. Lerat, F. Brunet, C. Bazin, and P. Capy. Is the evolution of transposable elements modular ?. *Genetica*, 107 :15–25, 1999.

[127] I.C. Lerman. Likehood linkage analysis (lla) classification method ; an example treated by hand. *Biochimie*, 75 :379–397, 1993.

[128] I.C. Lerman, P. Peter, and H. Leredde. Principes et calculs de la méthode implémenté dans le programme chavl (classification hiéarchique par analyse de la vraisemblance des liens). *Modulad*, pages 33–101, 1993.

[129] M. Lescot, P. Déhais, G. Thijs, K. Marchal, Y. Moreau, de. Van, P. Rouzé, and S. Rombauts. Plantcare and a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.*, 30 :325–327, 2002.

[130] S. Leung, C. Mellish, and D. Robertson. Basic gene grammars and dna-chartparser for language processing of escherichia coli promoter dna sequences. *Bioinformatics*, 17 :226–236, 2001.

[131] W. Li, P. Bernaola-Galván, F. Haghighi, and I. Grosse. Applications of recursive segmentation to the analysis of dna sequences. *Computers and Chemistry*, 26 :491–510, 2002.

[132] K. Liolios, N. Tavernarakis, P. Hugenholtz, and NC. Kyrpides. The genomes on line database (gold) v.2 : a monitor of genome projects worldwide. *Nucleic Acids Res.*, 34 :D332–D334, 2006.

[133] C. Loot, N. Santiago, A. Sanz, and J.M. Casacuberta. The proteins encoded by the pogo-like lemi1 element bind the tirs and subterminal repeated motifs of the arabidopsis emigrant mite : consequences for the transposition mechanism of mites. *Nucleic Acids Res.*, 34 :5238–5246, 2006.

[134] H.S. Malik. The rte class of non-ltr retrotransposons is widely distributed in animals and is the origin of many sines. *Mol. Biol. Evol.*, 15 :1123–1134, 1998.

[135] U. Manber and G. Myers. Suffix arrays : a new method for on-line string searches. *SIAM Journal on Computing*, 22 :935–948, 1993.

[136] L. Marsan and M.F. Sagot. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *JCB*, 7 :345–362, 2000.

[137] RM. Marsano, S. Marconi, R. Moschetti, P. Barsanti, C. Caggese, and R. Caizzi. Max and a novel retrotransposon of the bel-pao family and is nested within the bari1 cluster at the heterochromatic h39 region of chromosome 2 in drosophila melanogaster. *Mol. Genet. Genomics*, 270 :477–484, 2004.

[138] JM. Mason, AY. Konev, MD. Golubovsky, and H. Biessmann. Cis- and trans-acting influences on telomeric position effect in drosophila melanogaster detected with a subterminal transgene. *Genetics*, 163 :917–930, 2003.

[139] GA. Maston, SK. Evans, and MR. Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7 :29–59, 2006.

[140] V. Matys, E. Fricke, R. Geffers, E. Göbling, M Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E Kel, O.V Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Mnch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. Transfac : transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31 :374–378, 2003.

[141] E. McCarthy and J. McDonald. Ltr_struc : a novel search and identifica-tion program for ltr retrotransposons. *Bioinformatics*, 19 :362–367, 2003.

[142] B. McClintock. Chromosome organization and genic expression. *Cold Spring Harbor Symp. Quant. Biol.*, 16 :13–47, 1951.

[143] E.M. McCreight. A space-economical suffix tree construction algorithm. *Journal of the ACM.*, 23 :262–272, 1976.

[144] P. Medstrand and D.L. Mager. Human-specific integrations of the herv-k endogenous retrovirus family. *Journal of Virology*, 72 :9782–9787, 1998.

[145] M.V. Mendiola, I. Bernales, and la. De. Differential roles of the transposon termini in is91 transposition. *P.N.A.S.*, 91 :1922–1926, 1994.

[146] M.V. Mendiola, Y. Jubete, and la. De. Dna sequence of is91 and identification of the transposase gene. *J. Bact.*, 174 :1345–1351, 1992.

[147] C. Merlin and A. Toussaint. Les éléments transposables bactériens. *médecine/sciences*, 15, 1999.

[148] S. Moreno-Vazquez, J. Ning, and BC. Meyers. hatpin and a family of mite-like hat mobile elements conserved in diverse plant species that forms highly stable secondary structures. *Plant Mol Biol.*, 58 :869–886, 2005.

[149] M. Morgante, S. Brunner, G. Pea, K. Fengler, and A. Zuccolo. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genet.*, 37 :997–1002, 2005.

[150] B. Morgenstern, K. Frech, A. Dress, and T. Werner. Dialign : finding local similarities by multiple sequence alignment. *Bioinformatics*, 14 :290–294, 1998.

[151] T. Muller, M. Deschauer, F. Kolbe-Fehr, and S. Zierz. Genetic heterogeneity in 30 german patients with oculopharyngeal muscular dystrophy. *J Neurol.*, 253 :892–895, 2006.

[152] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5 :32–38, 1957.

[153] C. Mézard. Meiotic recombination hotspots in plants. *Biochem Soc Trans.*, 34 :531–534, 2006.

[154] J. Nicolas, P. Durand, G. Ranchy, S. Tempel, and AS. Valin. Suffix-tree analyser (stan) : looking for nucleotidic and peptidic patterns in chromosomes. *Bioinformatics*, 21 :4408–4410, 2005.

[155] T.R. O'Connor, C. Dyreson, and J.J. Wyrick. Athena : a resource for rapid visualization and systematic analysis of arabidopsis promoter sequences. *Bioinformatics*, doi :10.1093 :1–5, 2005.

[156] I. Ogiwara, M. Miya, K. Ohshima, and N. Okada. Retropositional parasitism of sines on lines : Identification of sines and lines in elasmobranchs. *Mol. Biol. Evol.*, 16 :1238–1250, 1999.

[157] J. Oliver, P. Carpena, M. Hackenberg, and P. Bernaola-Galván. Isofinder : computational prediction of isochores in genome sequences. *Nucleic Acids Research*, 32 :287–292, 2004.

[158] E.M. Ostertag. Twin priming : A proposed mechanism for the creation of inversions in l1 retrotransposition. *Genome Res.*, 11 :2059–2065, 2001.

[159] T. Palomeque, Carrillo. Antonio, M. Munoz-Lopez, and P. Lorite. Detection of a mariner-like element and a miniature inverted-repeat transposable element (mite) associated with the heterochromatin from ants of the genus messor and their possible involvement for satellite dna evolution. *Gene*, 371 :194–205, 2006.

[160] X. Pan, H. Liu, J. Clarke, J. Jones, M. Bevan, and L. Steina. Atidb : Arabidopsis thaliana insertion database. *Nucleic Acids Res.*, 31 :1245–1251, 2003.

[161] L. Peshkin and M. Gelfand. Segmentation of yeast dna using hidden markov models. *Bioinformatics*, 15 :880–986, 1999.

[162] A. Petit, F. Rouleux-Bonnin, M. Lambelé, N. Pollet, and Y. Bigot. Properties of the various botmar1 transcripts in imagoes of the bumble bee, bombus terrestris (hymenoptera :apidae). *Gene*, 390 :52–66, 2007.

[163] T. Pélissier, C. Bousquet-Antonelli, L. Lavie, and Deragon. and. Synthesis and processing of trna-related sine transcripts in arabidopsis thaliana. *Nucleic Acids Res.*, 32 :3957–3966, 2004.

[164] P. Polak and E. Domany. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics*, 7 :E133, 2006.

[165] S.S. Potter. Dna sequence of a foldback transposable element in drosophila. *Nature*, 297 :201–204, 1982.

[166] R.T.M. Poulter, T.J.D. Goodwin, and M.I. Butler. Vertebrate helentrons and other novel helitrons. *Gene*, 313 :201–212, 2003.

[167] AL. Price, NC. Jones, and PA. Pevzner. De novo identification of repeat families in large genomes. *Bioinformatics*, Suppl 1 :i351–i358, 2005.

[168] E. Prieur and T. Lecroq. From suffix trees to suffix vectors. *In Proceedings of PSC'05 and Prague and Czech Republic*, 2005.

[169] E.J. Pritham and C. Feschotte. Massive amplification of rolling-circle transposons in the lineage of the bat myotis lucifugus. *P.N.A.S.*, 104 :1895–1900, 2007.

[170] H. Quesneville, CM. Bergman, O. Andrieu, D. Autard, D. Nouaud, M. Ashburner, and D. Anxolabehere. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol.*, 1 :166–175, 2005.

[171] D. Reiss, T. Josse, D. Anxolabehere, and S. Ronsseray. Aubergine mutations in drosophila melanogaster impair p cytotype determination by telomeric p elements inserted in heterochromatin. *Mol. Genet. Genomics.*, 272 :336–343, 2004.

[172] SY. Rhee, W. Beavis, TZ. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, LA. Mueller, S. Mundodi, L. Reiser, J. Tacklind, DC. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang. The arabidopsis information resource (tair) : a model organism database providing a centralized and curated gateway to arabidopsis biology and research materials and community. *Nucleic Acids Res.*, 31 :224–228, 2003.

[173] M. Rho, J.H. Choi, S. Kim, M . Lynch, and H. Tang. De novo identification of ltr retrotransposons in eukaryotic genome. *BMC Genomics*, 8 :90–125, 2007.

[174] HM. Robertson. The mariner transposable element is widespread in insects. *Nature*, 362 :241–245, 1993.

[175] S. Rombauts, K. Florquin, M. Lescot, K. Marchal, P. Rouzé, and Van. and. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiology*, 132, 2003.

[176] F. Sabot and D. Simon. Plant transposable elements and with an emphasis on grass species. *Euphytica*, 139 :227–247, 2004.

[177] N. Saitou and M. Nei. The neighbour-joining method : a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4 :406–425, 1987.

[178] J. Salse, B. Piégu, R. Cooke, and M. Delseny. Synteny between arabidopsis thaliana and rice at the genome level : a tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res.*, 30 :2316–2328, 2002.

[179] D. Samuels, R. Boys, D. Henderson, and PF. Chinnery. A compositional segmentation of the human mitochondrial genome is related to heterogeneities in the guanine mutation rate. *Nucleic Acids Res.*, 31 :6043–6052, 2003.

[180] F. Sanger, GM. Air, BG. Barrell, NL. Brown, AR. Coulson, CA. Fiddes, CA. Hutchison, PM. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi x174 dna. *Nature*, 265 :687–695, 1977.

[181] D. Searls. Investigating the linguistics of dna with definite clause grammars. *In Lusk and E. and Overbeek and R. (eds) and Logic programming : Proceedings of the North American Conference on Logic Programming. MIT Press*, pages 189–208, 1989.

[182] D.B. Searls. String variable grammar : a logic grammar formalism for the biological language of dna. *Journal of logic programming*, 12 :1–30, 1993.

[183] D.B. Searls. The language of genes. *Nature*, 420 :211–217, 2002.

[184] D.B. Searls and S. Dong. Gene structure prediction by linguistic methods. *Genomics*, 23 :540–551, 1994.

[185] F. Servant, C. Bru, S. Carrere, E. Courcelle, J. Gouzy, D. Peyruc, and D. Kahn. Prodom : automated clustering of homologous domains. *Brief Bioinform.*, 3 :246–251, 2002.

[186] R. Shankar, D. Grover, SK. Brahmachari, and M. Mukerji. Evolution and distribution of rna polymerase ii regulatory sites from rna polymerase iii dependant mobile alu elements. *BMC Evol Biol.*, 4, 2004.

[187] T. Singer, C. Yordan, and Martienssen. and. Robertson s mutator transposons in a. thaliana are regulated by the chromatin-remodeling gene decrease in dna methylation (ddm1). *Genes and Dev.*, 15 :591–602, 2001.

[188] L. Sinzelle, A. Chesneau, Y. Bigot, A. Mazabraud, and N. Pollet. The mariner transposons belonging to the irritans subfamily were maintained in chordate genomes by vertical transmission. *J. Mol. Bio.*, 62 :53–65, 2006.

[189] C.H. Slamovits and M.S. Rossi. Satellite dna : agent of chromosomal evolution in mammals. a review. *J. Neotrop. Mammal.*, 9 :297–308, 2002.

[190] E.E. Slawson, C.D. Shaffer, C.D. Malone, W. Leung, E. Kellmann, R.H. Shevchek, C.A. Craig, S.A. Bloom, J. Bogenpohl, J. Dee, E.TA. Morimoto, J. Myoung, A.S. Nett, F. Ozsolak, M.E. Tittiger, A. Zeug, M. Pardue, J. Buhler, E.R. Mardis, and S.CR. Elgin. Comparison of dot chromosome sequences from d. melanogaster and d. virilis reveals an enrichment of dna transposon sequences in heterochromatic domains. *Genome Biol.*, 7 :R15, 2006.

[191] C.D. Smith, R.C. Edgar, M.D. Yandell, D.R. Smith, S.E. Celniker, E.W. Myers, and G.H. Karpen. Improved repeat identification and masking in dipterans. *Gene*, 389 :1–9, 2007.

[192] TF. Smith and M.S. Waterman. Identification of common molecular subsequences. *J Mol Biol.*, 147 :195–197, 1981.

[193] T. Strachan and A. Read. Human molecular genetics. 3rd edition. *ISBN-13 : 978-0-8153-4182-6*, 2003.

[194] S. Subramanian, R.K. Mishra, and L. Singh. Genome-wide analysis of microsatellite repeats in humans : their abundance and density in specific genomic regions. *Genome Biology*, 4 :R13, 2003.

[195] S.A. Surzycki. Characterization of repetitive of arabidopsis thaliana dna elements in arabidopsis. *J. Mol. Evol.*, 48 :684–91, 1999.

[196] S. Tempel, I.C. Lerman, M. Giraud, A.S. Valin, I. Couée, Amrani. El, and J. Nicolas. Domain organization within repeated dna sequences : application to the study of a family of transposable elements. *Bioinformatics*, 22 :1948–54, 2006.

[197] European. The, University. Washington, and Biosystems. PE. Sequence and analysis of chromosome 4 of the plant arabidopsis thaliana. *Nature*, 14 :769–777, 2000.

[198] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, Moor. De, P. Rouzé, and Y. Moreau. A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9 :447–464, 2002.

[199] G. Thijs, Y. Moreau, F. De Smet, J. Mathys, M. Lescot, S. Rombauts, P. Rouzé, B. De Moor, and K. Marchal. Inclusive : Integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, 18 :331–332, 2002.

[200] J.D. Thompson, . Higgins, . D.G, . Gibson, and . T.J.

[201] BG. Thornburg, V. Gotea, and W. Makalowski. Transposable elements as a significant source of transcription regulating signals. *Gene*, 365 :104–10, 2006.

[202] A. Toussaint and C. Merlin. Mobile elements as a combination of functional modules. *Plasmid*, 47 :26–35, 2002.

[203] E. Ukkonen. On-line construction of suffix-trees. *Algorithmica*, 14 :249–260, 1995.

[204] J. van Helden. Regulatory sequence analysis tools. *Nucleic Acids Res*, 31 :3593–3596, 2003.

[205] N. Volfovsky, B.J. Haas, and S. Salzberg. A clustering method for repeat analysis in dna sequences. *Genome Biology*, 2 :1–11, 2001.

[206] W. Wei and MD. Brennan. The gypsy insulator can act as a promoter-specific transcriptional stimulator. *Mol. Cell Biol.*, 21 :7714–20, 2001.

[207] P. Weiner. Linear pattern matching algorithm. *14th Annual IEEE Symposium on Switching and Automata Theory*, pages 1–11, 1973.

[208] T. Wicker, R. Guyot, N. Yahiaoui, and Keller. and. Cacta transposons in triticeae. a diverse family of high-copy repetitive elements. *Plant Physiol.*, 132 :52–63, 2003.

[209] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer. Transfac : an integrated system for gene expression regulation. *Nucleic Acids Res*, 28 :316–9, 2000.

[210] J.H. Xu and J. Messing. Maize haplotype with a helitron-amplified cytidine deaminase gene copy. *BMC Genetics*, 7 :1–13, 2006.

[211] D. Zhi, B.J. Raphael, A.L. Price, H. Tang, and P.A. Pevzner. Identifying repeat domains in large genomes. *Genome Biology*, 7, 2006.

[212] P. Zimmermann, M. Hirsch-Hoffmann, L. Hennig, and W. Gruissem. Genevestigator. arabidopsis microarray database and analysis toolbox. *Plant Physiol.*, 136 :2621–2632, 2004.

[213] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31 :3406–15, 2003.

Table des figures

Résumé

Les hélitrons constituent un groupe d'éléments transposables découverts récemment dans les génome eucaryotes. A travers une étude bioinformatique, nous avons étudié leur mode d'invasion, la modularité de leur séquence et leurs impacts sur les gènes à leur proximité dans le génome d'*Arabidopsis thaliana*.

Les hélitrons sont les éléments transposables les plus répandus dans ce génome ; néanmoins ils ne sont que partiellement reconnus par des logiciels d'alignement. Nous avons modélisé ces éléments sous la forme d'une grammaire formelle. Cette grammaire est constituée des deux extrémités terminales séparées par une séquence nucléotidique quelconque de taille fixée. Nous avons créé une matrice d'occurrences des modèles associant toutes les combinaisons possibles d'extrémités. La matrice a fait apparaître des associations préférentielles entre certaines extrémités et a permis la découverte de nouvelles familles d'hélitrons chimériques. La détection des ORFs contenant les protéines de transposition a permis de confirmer la relation hélitron autonome non-autonome et de comprendre le mécanisme de création des chimères d'hélitrons. Nous avons proposé une nouvelle nomenclature des hélitrons basée sur leurs extrémités et non sur leur séquence globale.

L'étude de la séquence d'une famille d'hélitrons a montré une réorganisation constante des domaines nucléiques entre les différentes copies de cette famille. Pour comprendre cette organisation, nous avons mis au point le logiciel DomainOrganizer qui permet d'observer la composition en domaines des éléments transposables. DomainOrganizer détecte les frontières entre domaines à partir d'un alignement multiple et crée la liste des domaines. A partir de cette liste, il recherche, par un algorithme d'optimisation combinatoire, le nombre minimal de domaines qui recouvrent au maximum l'ensemble des séquences. Enfin, DomainOrganizer visualise et classe les séquences en fonction de leurs domaines. L'analyse par domaines de la famille AtREP21 a permis de comprendre la nature de cette variabilité et de retracer l'histoire évolutive de cette famille à partir de l'identification des domaines.

L'étude de la localisation des hélitrons AtREP3 dans ce génome de plante a montré une insertion préférentielle de ceux-ci dans les promoteurs de gènes. Les profils d'expression de ces gènes, nous a permis d'identifier plusieurs clusters. Par ailleurs, les motifs de régulation ont montré une grande variabilité de motifs dans les promoteurs mais pas dans les hélitrons. Ces résultats ont montré que les hélitrons non-autonomes transportent dans leurs séquences internes des motifs de liaisons aux facteurs de transcription. Des analyses complémentaires devront être réalisées pour comprendre l'action régulatrice des hélitrons sur les gènes situés à leur proximité.

Mots Clés : Arbres des suffixes, *Arabidopsis thaliana*, Chimère, Co-régulation, Domaine nucléique, Elément transposable, Hélitron, Modèle syntaxique, Optimisation combinatoire.