Création par évolution dirigée de protéines artificielles en alternatives aux anticorps

Asma Guellouz

▶ **To cite this version:**

Asma Guellouz. Création par évolution dirigée de protéines artificielles en alternatives aux anticorps. Biochimie [q-bio.BM]. Université Paris Sud - Paris XI, 2012. Français. NNT : 2012PA11T063 . tel-00767675

HAL Id: tel-00767675

https://theses.hal.science/tel-00767675

Submitted on 20 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Paris Sud
La Faculté de Médecine Paris-Sud

Ecole doctorale :

Signalisations et Réseaux Intégratifs en Biologie BIO-SigNE

Année 2012-2013 DOCTORAT EN BIOCHIMIE N°2012 PA11 T063

Thèse de doctorat présentée

À l'Institut de Biochimie et Biophysique Moléculaire et Cellulaire

Pour l'obtention du grade de

Docteur de l'Université Paris Sud 11

Par

Mlle Asma Guellouz

أسماء قلوز

Titre de la thèse

Création par évolution dirigée de protéines artificielles en alternative aux anticorps.

Soutenue le 25/10/2012

Directeur de thèse :	***Philippe Minard***	PU1, Université Paris Sud 11
Composition du Jury :		
Rapporteurs :	***Pierre Martineau***	CR1, IRC Montpellier.
	Charles Tellier	PU, Université de Nantes.
Examinateurs :	***Cécile Van De Weerdt***	Chef de projet, GIGA, Belgique.
	Franc Perez	DR2, Institut Curie, Paris.
Président du jury :	***Herman Van Tilbeurgh***	PU1, Université Paris Sud 11.

Je souhaite dédier ma thèse à toutes les personnes qui m'ont aidée ou soutenue au cours de ces dernières années et une pensée particulière à

A mes parents Halima et khaled, et mes sœurs

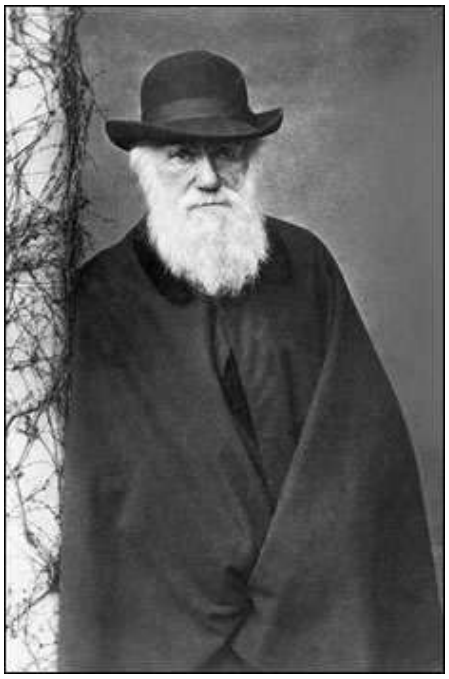
A Mes ami (e)s……

Je souhaite remercier tous les membres de mon équipe ; Agathe, Marielle, Magalie, Anne et Philippe ainsi que toutes les équipes de l'IBBMC qui ont contribué de près ou de loin à la réalisation de ces travaux en particulier l'équipe Fonction et Architecture des Assemblages Macromoléculaires (FAAM).

Je souhaite également remercier l'IBBMC et la fondation de la recherche médicale pour leur soutien financier pendant ma thèse.

«Les espèces qui survivent ne sont pas les espèces les plus fortes, ni les plus intelligentes, mais celles qui s'adaptent le mieux aux changements.»

Charles Darwin



Sommaire général

Introduction générale

Chapitre I: Conception et création d'une banque de protéines artificielles : les *αRep*

Chapitre II : Création d'une banque d'*αRep* de deuxième génération2.1, sélection d'interacteurs pour des cibles choisies et caractérisation des interactions

A. Optimisation et construction du vecteur d'expression

B. Optimisation de la fraction des clones codants dans la banque

C. Construction de la banque d'αRep de deuxième génération 2.1

D. Sélection, identification d'interacteurs pour des molécules cibles et caractérisation des complexes obtenus

E. Travaux supplémentaires

Conclusion et perspectives

Introduction générale

Table des matières

Abréviations

ADLib: *Autonomously Diversifying Library*

ADN : Acide Désoxyribonucléique

ADNc : ADN complémentaire

AID: la Cytidine Désaminase Inductible

ARN : Acide Ribonucléique

ARNm : ARN messagers

ARNt : ARN de transfert

BFP: *Blue Fluorescent Protein*

BSA: *Bovine Serum albumin*

Chaînes H et L : chaînes *Heavy* (lourde) et *Light* (légère)

cfu : *Colony forming unit*

DARPin : *Designing Ankyrin Repeat Proteins*

DHFR : Dihydrofolate Réductase

dNTP : mélange des quatre désoxyribonucléotides (dATP (désoxy adénine tri-phosphate), dCTP (désoxy cytosine tri-phosphate), dGTP (désoxy guanine tri-phosphate), dTTP (désoxy thymine tri-phosphate)

DuARCheM : *Dual Approach to Random Chemical Mutagenesis*

DXP: 1desoxy-D-xylulose-5-phosphate

EETI-II: *Ecballium elaterium tripsin inhibitor*

EMS : Ethylméthane sulfonate

EPO : érythropoïétine

Fab : *Fragment antigen binding*

FACS : *Fluorescence-activated cell sorting*

FDA: *U S Food and Drug Administration*

GFP : *Green Fluorescent Protein*

HEK : *Human Embryonic Kidney*

HER2: *Human Epidermal Growth Factor Receptor-2*

IgG : Immunoglobulines du groupe G

Il-6R: Interleukine 6

IPTG: IsoPropyl β-D-1-ThioGalactopyranoside

k_D : Constante de dissociation

MAGE: *Multiplex Automated Genome Engineering*

mRFP: *monomeric Red Fluorescent Protein*

LRR*: Leucines Rich repeat*

PCA : *Protein-fragment Complementation Assay*

PCR : *Polymerase Chain reaction* (Réaction en chaîne par polymérase)

PP5: Phosphatase 5

RCA : *Rolling Circle Amplification*

RT-PCR : Reverse Transcriptase PCR

scFv : *single-chain variable fragment*

SELEX: *Systematic Evolution of Ligands by EXponential Enrichement*

SRP: *Signal Recognition Particle*

SNP : *Single Nucleotide Polymorphisms*

SSM : *Site Saturation Mutagenesis*

STABLE : *STA-Biotin Linkage in Emulsions*

TNF: *Tumor Necrosis Factor* (Facteur de nécrose tumorale)

TPR: *Tetratrico Peptide Repeat*

UAA : *Unnatural Amino Acids*

VEGFR-2: *Vascular Endothelial Growth Factor (VEGF) Receptor 2*

VLR: *Variable Lymphocyte Receptors*

Introduction

Toutes les fonctions protéiques, aussi diverses soient-elles, résultent en partie de leur étonnante capacité de reconnaissance moléculaire. En effet qu'il s'agisse de catalyse, de régulation, de transport ou de communication, la fonction d'une protéine suppose qu'elle établisse des interactions spécifiques avec un ou plusieurs partenaires moléculaires. Le répertoire des interactions naturellement exercées par des protéines apparaît donc comme très riche par la diversité des objets reconnus et par la variété des processus biologiques que cette reconnaissance permet. Parvenir à recréer des protéines ayant des capacités d'interactions choisies apparait donc comme un questionnement fondamental sur l'émergence des propriétés des protéines naturelles. C'est également une étape qui devra être franchie pour parvenir à créer des nouvelles protéines capables d'exercer des fonctions nouvelles et utiles.

Pour atteindre cet objectif, deux grandes familles d'approches ont été envisagées : il s'agit d'une part des approches de conception rationnelle de protéine ou *« Protein Design* » et d'autre part des approches d'évolution dirigée d'ossatures protéiques naturelles.

A. Conception rationnelle ou « Protein Design »

La conception rationnelle de protéines ou *« Protein Design »* est une approche qui consiste à définir une structure tridimensionnelle protéique susceptible d'établir les interactions recherchées, puis à concevoir une séquence polypeptidique susceptible de se replier dans cette conformation (Baker, 2006). En partie grâce à l'accumulation des données structurales, cette approche a effectué de grands progrès, tout d'abord dans la résolution de problèmes « simples » tels que la prédiction de l'effet d'une mutation unique sur la stabilité d'une protéine ou sur l'affinité d'une interaction. La résolution de problèmes plus compliqués, tels que la prédiction des changements de structure résultant de l'accumulation de mutations non-conservatives, demeure difficile et couteuse en temps de calcul (Verschueren et *al.*, 2011). Un autre défi a aussi suscité l'intérêt des chercheurs : prédire un ensemble de séquences tolérées par une protéine ou par une interface d'interaction protéine-protéine tout en conservant la fonction désirée. Ainsi on peut caractériser des interactions et concevoir des banques de protéines dotées de nouvelles fonctions (C. A.Smith and Kortemme, 2011). Cette méthode a permis plusieurs autres applications telles que la stabilisation des interactions avec d'autres molécules (ADN ou ARN).

L'équipe de *D. Baker* a mis au point des méthodes innovantes permettant la conception de nouvelles structures protéiques, la conception de nouvelles interactions protéine-protéine, et la conception de sites catalytiques nouveaux capables de réaliser des réactions enzymatiques.[1]

Malgré ces succès remarquables, ces approches ne sont pas diffusées largement et, de par leur difficulté, ne sont pas encore l'option la plus réaliste pour résoudre de nombreux problèmes concrets. Par exemple, la conception de nouvelles protéines capables d'interagir avec une molécule cible donnée ; protéine ou acide nucléique a été réalisée par des approches théoriques, mais cela demeure à chaque fois un problème nouveau et difficile.

En effet nos connaissances des règles reliant séquence et structure demeurent imprécises pour arriver à concevoir *in silico* une séquence qui donnera une structure donnée ou encore la structure tridimensionnelle nécessaire pour l'interaction recherchée. C'est dans le but de contourner les difficultés des approches de conception rationnelle, que des approches d'évolution dirigée des protéines ont été mises en place.

[1] http://depts.washington.edu/bakerpg/drupal/node/10

B. Evolution dirigée :

1. Définition

L'évolution dirigée[2] est un ensemble de technologies permettant d'améliorer une protéine ou un acide nucléique, en reproduisant artificiellement le processus naturel de l'évolution mais en cherchant à l'orienter dans une direction choisie. Les approches d'évolution dirigée nécessitent deux étapes essentielles :

- une étape de création de la diversité génétique
- une étape de criblage ou de sélection des variants présentant des propriétés recherchées

La première étape de création de diversité consiste à créer à partir d'un gène initial une population de gènes codants pour des formes modifiées de la protéine d'intérêt. Ces modifications peuvent être totalement aléatoires par leur position et leur nature ou bien aléatoires mais localisées à des positions particulières. Elles peuvent aussi résulter d'un processus de recombinaison de séquences préalablement modifiées. On appelle banque, ou bibliothèque, la population ainsi obtenue. Les grandes étapes ayant permis le développement de ces technologies sont exposées dans les paragraphes suivants.

2. Les technologies recombinantes

C'est un ensemble de connaissances fondamentales qui a permis la mise au point d'outils nécessaires aux développements des technologies de modification de l'ADN. La découverte fondatrice est sans doute l'isolement des endonucléases de restriction. Tout d'abord la découverte et la caractérisation de HindIII en 1970 et par la suite d'autres enzymes de restriction, ont valu le prix Nobel de Médecine et de Physiologie à Daniel Nathans, Werner Arber, et Hamilton O. Smith en 1978. Ces enzymes, mises en évidence chez les bactéries ont évoluées formant un mécanisme de défense contre les envahisseurs (Krüger DH and Krüger DH, Bickle TA, 1983). Cette découverte importante suivie par l'utilisation d'autres enzymes de modification de l'ADN, ligases et polymérases, a conduit au développement des techniques de recombinaison de l'ADN. Par la suite, il y a eu le développement des méthodes de séquençage et les techniques de synthèse d'ADN soit par oligonucléotides (Caruthers et

[2]Thèse de doctorat d'Antoine Drevelle : Evolution Dirigée de la néocarzinostatine : ingénierie d'une ossature protéique alternative aux anticorps. Université Paris-Sud11 année 2006-2007.

al., 1980) soit avec l'invention de la PCR par *Kary Mullis* en 1983 (Bartlett and Stirling, 2003) en parallèle avec la diffusion des méthodes de mutagénèse dirigée par oligonucléotides.

Les techniques de génie génétique ont trouvé des applications concrètes et variées, et en particulier des applications médicales. Des protéines telles que l'insuline, l'hormone de croissance humaine et les anticorps sont désormais produites dans des bactéries ou des cellules eucaryotes et grâce au développement parallèle des méthodes de purification et de caractérisation des protéines, elles peuvent être utilisées en thérapeutique humaine.

L'ingénierie des protéines a émergé grâce à la manipulation génétique et les technologies recombinantes mais aussi au développement de la biochimie structurale pour l'analyse et la compréhension des conséquences des modifications de séquences sur la structure des protéines.

3. L'évolution moléculaire dirigée

Grâce aux outils de la biologie moléculaire, il est possible de mimer en laboratoire les processus à l'origine de l'évolution naturelle, ce qui a permis de modifier les propriétés des molécules biologiques telles que les protéines, ce qui aurait été sans doute impossible à obtenir autrement.

Les deux composantes de l'évolution naturelle sont reproduites artificiellement : la génération de la diversité est obtenue en agissant sur l'ADN (génotype) et la population est conduite artificiellement à évoluer par la sélection des mutants ayant intégré le caractère d'intérêt (phénotype). L'objectif essentiel de ce processus est de parvenir à générer des mutants ayant la caractéristique recherchée et puis de parvenir à isoler l'information génétique correspondante et à l'amplifier.

La toute première protéine qui a été sujette à l'évolution dirigée est une forme de β-galactosidase. A côté de la β-galactosidase classiquement étudiée et codée par le gène lacZ sur l'opéron lactose, il existe chez *E. coli* une autre protéine modèle possédant une activité β-galactosidase. En effet, l'opéron de l'*Ebg* d'*E.coli* a été utilisé depuis 1974 comme un modèle pour l'étude des processus d'évolution de l'efficacité catalytique et la spécificité des substrats chez les enzymes (Hall, 1999). Cette β-galactosidase, codée par le gène *ebgA*, est une enzyme à faible activité catalytique ne permettant pas une croissance bactérienne sur

certains types de substrats (lactose…). Ceci a permis de l'utiliser comme modèle pour l'étude de l'évolution de nouvelles fonctions car :

- une forte pression de sélection peut être appliquée à une large population permettant de révéler des mutations spontanées rendant le métabolisme du lactose possible.
- une grande précision est possible dans la manipulation des conditions sélectives.
- les 2 états, sauvage et mutant, seront par la suite caractérisés et comparés dans le but de déterminer les changements génotypiques responsables de l'évolution du phénotype Lac^+.

En effet, l'isolement, à partir de souche Δlac*Z* *ebgR* (constitutive), de mutants qui croissent sur un milieu riche en lactose a montré l'apparition de mutations (substitutions) dans le gène *ebgA*. Des travaux ultérieurs combinant les 2 mutations identifiées ont permis d'augmenter encore plus l'efficacité catalytique et d'élargir la gamme de substrats de l'enzyme.

Ces résultats prometteurs ont été précurseurs pour les travaux d'évolution dirigée des protéines. Ces méthodes ont ensuite considérablement progressé avec le développement des méthodes de génération de diversité et des méthodes de criblage et de sélections *in vitro* et *in vivo*. Ces techniques seront décrites tout au long de ce chapitre.

4. Les techniques d'évolution moléculaire dirigée des protéines

Quelle méthode pour quel objectif ?

La réussite du processus d'évolution réside dans la bonne combinaison entre les méthodes de génération de diversité et celles qui permettent de sélectionner ou de cribler les mutants ayant incorporés la propriété voulue. Les deux aspects, diversification et sélection, doivent être adaptés l'un à l'autre mais doivent également être adaptés à la transition évolutive qui est recherchée. On peut schématiquement distinguer deux types de situations selon que l'objectif soit de créer une nouvelle fonction ou qu'il soit d'améliorer une ou des propriétés préexistantes d'une protéine. On conçoit bien que, s'il s'agit d'améliorer l'affinité et/ou la spécificité d'un anticorps pour un antigène, ou encore d'accroître l'activité catalytique ou la stabilité d'une enzyme, un processus global basé sur une série d'adaptations

progressives sera possible et une amélioration globale du phénotype provenant d'un cumul de petits pas évolutifs parait la stratégie la plus sûre.

Le problème est très différent dans le second type de situation, lorsqu'il s'agit non seulement d'améliorer mais de créer: la création d'une interface nouvelle, ou d'une activité catalytique, qui ne préexiste pas dans la protéine initiale, ne pourra pas être obtenue par des améliorations progressives, et c'est un saut évolutif plus important qui doit être d'emblée réalisé.

Quelle est la part du hasard ?

Plusieurs méthodes ont été développées dans le but de construire des banques de protéines telles que la substitution aléatoire, le brassage d'ADN (DNA *Shuffling)*… Ces stratégies introduisent des variations en tout point de la séquence d'une protéine naturelle et permettent d'obtenir des bibliothèques sans tenir compte des données structurales des protéines d'intérêt. Mais est-il absolument nécessaire de partir d'une protéine naturelle ? Une forme encore plus extrême d'évolution dirigée des protéines consiste à rechercher l'émergence d'une fonction biologique à partir de séquences polypeptidiques complétement aléatoires [(Matsuura and Yomo, 2006) ; (Keefe and Szostak 2001)**]**. La difficulté réside dans le fait qu'il est fort probable que même parmi une collection diversifiée de protéines, il n'existe aucune séquence capable de se replier de façon stable, dans la conformation requise pour la propriété recherchée. Les probabilités de repliement et d'adaptation fonctionnelle bien que paraissant intuitivement très faibles, ne peuvent pourtant être nulles, puisque les protéines naturelles ont nécessairement émergé à partir de séquences aléatoires. Les résultats obtenus dans quelques cas montrent que la recherche de fonctions, à partir de protéines totalement aléatoires, peut effectivement aboutir. Pour autant, les travaux de ce type demeurent peu nombreux. Une démarche totalement aléatoire dans l'espace des séquences présente sans doute un risque élevé et suppose, au minimum, d'explorer des bibliothèques exceptionnellement diverses. Il parait plus réaliste, de focaliser l'exploration sur une population de séquences ayant une probabilité plus élevée d'avoir une structure tridimensionnelle. D'où la nécessité de créer des banques de protéines qui ne soient pas complètement aléatoires mais plutôt des bibliothèques combinatoires basées sur des architectures protéiques naturelles. Ces bibliothèques se composent de variants de protéines naturelles qui ne sont mutés aléatoirement que sur une partie de leur séquence comme par exemple sur une ou plusieurs boucles ou sur des résidus de surface composant un site

d'interaction avec les ligands. Cette démarche vise à garder une partie de la séquence capable de se replier en préservant une structure nécessaire pour exposer la surface d'interaction potentielle.

L'essentiel des travaux réalisés partent donc d'une structure ou d'une activité biologique existante. Il s'agit de générer des populations de variants de cet objet initial, puis de sélectionner. Les outils de modification, d'optimisation de protéines et de création de nouvelles fonctionnalités, peuvent être divisés en deux catégories différentes : des techniques purement *in vitro* où les 2 étapes sont réalisées *in vitro* et d'autres où la mutagenèse est effectuée *in vitro* et le criblage nécessite une étape de transformation ou de transfection *in vivo*. Les techniques les plus communément utilisées sont celles basées sur les méthodes *in vitro* telles que la mutagenèse aléatoire ou ciblée et la méthode de « *DNA shuffling* » pour la création de la diversité et le *ribosome ou le phage display* pour la sélection.

4.1. Techniques de création de la diversité *in vitro* :

Pour mimer l'évolution *in vitro*, le processus naturel doit être accéléré pour que la diversité soit créée puis criblée/sélectionnée dans un temps compatible avec les exigences expérimentales (Labrou, 2010). Les méthodes de génération de diversité les plus communément utilisées sont la mutagenèse aléatoire et la recombinaison. A titre d'exemple on peut citer : le traitement de l'ADN ou de la bactérie par des agents mutagènes chimiques, la PCR mutagène (*error-prone PCR*, *epPCR*), la RCA mutagène (*rolling circle error-prone* PCR, *epRCA*) et la mutagénèse à saturation. Le choix d'une stratégie ou d'une autre pour atteindre un but choisi n'est pas soumis à des règles précises, mais chaque méthode comporte des avantages particuliers et des limites.

a) Mutagénèse chimique :

Cette méthode implique l'utilisation d'agents chimiques tels que l'Ethylméthane sulfonate (EMS) qui agit sur les bases guanines ce qui induit des erreurs lors de la réplication de l'ADN. L'acide nitrique (HNO_2) est un autre agent qui agit en désaminant les adénines et les cytosines causant une mutation ponctuelle par transversion (A/T en G/C). Plusieurs autres agents ont été identifiés ainsi que leurs effets. Le bisulfite a été utilisé par l'équipe d'Ermakova-Gerdes pour muter aléatoirement et *in vitro* le gène *psbDI*. Le bisulfite de sodium agit spécifiquement sur la cytosine d'une région d'ADN simple brin. Dans cet

exemple, un plasmide hybride a été conçu, avec une région simple brin correspondant à la partie du gène psbDI à muter (codant pour la loupe A-B de la protéine D2) et une région double brin. Ainsi plusieurs variants ont été isolés puis analysés et au total 15 acides aminés différents ont été modifiés dans cette région (Ermakova-Gerdes et *al*., 1996).

Une nouvelle approche de mutagenèse a été présentée en 2008, DuARCheM (Dual Approach to Random Chemical Mutagenesis) permettant l'introduction de mutations aléatoires dans un fragment d'ADN bien défini (Mohan and Banerjee, 2008). Cette approche comporte une étape de mutagenèse chimique *in vivo* et des manipulations génétiques *in vitro*. La bibliothèque de mutants obtenue montre qu'il y a eu accumulation des mutations au cours des générations. De plus, cette méthode procure un meilleur spectre de mutants car elle permet non seulement d'introduire des mutations dans des régions ciblées mais aussi d'amplifier exponentiellement le matériel génétique (plasmide). Ceci permet de contourner les inconvénients associés aux mutagenèses chimiques aléatoires.

L'espace des séquences protéiques n'est plus limité au code génétique canonique, et une grande variété d'acides aminés non naturels (UAAs pour **U**nnatural **A**mino **A**cids) peut être incorporée dans les protéines. En effet, les UAAs peuvent être incorporés par remplacement des acides aminés naturels, lors de la traduction *in vitro* par le biais d'ARNt chimiquement aminoacétylés, par leur injection directe dans les cellules ou encore par génération d'ARNt / aminoacyl-ARNt synthétase pour l'incorporation spécifique d'acides aminés *in vivo*. L'incorporation d'UAAs dans les protéines a pour objectif de leur conférer de nouvelles propriétés chimiques exploitables pour des activités intéressantes. Par exemple, l'acide aminés coumarine fluorescent (codé génétiquement) a été introduit dans la région de liaison d'un Fab spécifique pour le CD40 avec son antigène. Les changements de fluorescence ont été utilisés pour contrôler la fixation du ligand, et fournir un moyen général pour la conception de biocapteurs basés sur des anticorps (Brustad and Arnold, 2011).

b) « Error Prone » PCR : epPCR :

Cette technique est basée sur une amplification par PCR de la séquence à modifier mais dans des conditions favorisant l'augmentation de la fréquence globale des erreurs de la DNA polymérase. Il est donc important d'utiliser une polymérase dépourvue de l'activité 3'- 5' exonucléase correctrice. Le taux d'erreur peut atteindre $0.8\text{-}1.1*10^{-4}$ substitutions/pb dans des conditions standards et peut être augmenté par l'ajout de Mn^{2+} (diminution de

l'efficacité de l'hybridation des bases) ou par l'augmentation de la concentration en Mg^{2+} dans le but de stabiliser les paires de bases non complémentaires. La fréquence des erreurs de polymérisation peut également être volontairement augmentée par l'utilisation d'un mélange de dNTP non équilibré induisant ainsi des non-appariements ou encore par l'augmentation de la concentration en polymérase afin d'augmenter la probabilité d'élongation des terminaisons sans amorces (Labrou, 2010). Après la PCR, la population des séquences modifiées doit être insérée dans un plasmide de façon à obtenir un ensemble de clones constituant la bibliothèque. La diversité des bibliothèques obtenues est habituellement limitée par l'efficacité de l'étape de ligation et de transformation plutôt que par la PCR elle-même. Dans l'*epPCR*, les mutations ponctuelles sont les plus fréquentes mais des délétions et donc des décalages du cadre de lecture peuvent aussi avoir lieu. Un exemple très récent, impliquant cette technique, est l'ingénierie de la sous-unité HycE de l'hydrogénase-3 d'*E.coli* pour laquelle, 7 variants améliorés ont été obtenus. La production d'hydrogène obtenue par la meilleure séquence est multipliée 17 fois relativement à celle obtenue avec la protéine initiale. Une analyse de séquence a mis en évidence que 8 mutations sont impliquées (Maeda et *al.*, 2008) dans cette amélioration.

Une autre procédure, MutaGen, présentée par l'équipe de Remaud-Simeon (Emond et *al.*, 2008) est basée sur l'utilisation des ADN polymérases Polβ et Polη (eta) à très faible fidélité. Cette technique a été appliquée pour muter un gène codant l'amylosucrase. Les auteurs ont déterminé que les mutants obtenus par l'ADN polymérase Polβ sont 4 à 7 fois moins mutés que ceux obtenus par la polymérase Polη, et que les spectres de mutations obtenus par l'une et l'autre sont complémentaires. La polymérase Polη a engendré des modifications comprenant des substitutions de bases et des délétions de codons, ce qui est rarement obtenu par d'autres méthodes.

Une stratégie aussi intéressante, intitulée « *frame shuffling* » a été décrite par l'équipe de Shiba (Kashiwagi et *al.*, 2006). Leurs travaux ont impliqué l'utilisation de l'ADN polymérase de famille Y qui introduit des taux importants de mutations « *frame shift* » pour changer les cadres de lectures des gènes parentaux. Les mutants produits montrent des échanges de segments de séquences et des propriétés physicochimiques différentes de celles des mutants obtenus par des méthodes de mutations conventionnelles.

c) Rolling Circle Amplification: RCA et epRCA :

Cette technique d'amplification circulaire (ou RCA pour Rolling Circle Amplification) est une méthode d'amplification de séquences *in vitro* adaptée à partir de la réplication d'ADN circulaire existant chez plusieurs virus (Fire and Xu, 1995). L'amplification, dans cette méthode, résulte d'une forte activité de déplacement de la Φ29 ADN polymérase. Cette méthode est isotherme car, contrairement à la PCR, elle est réalisée à température constante. Si l'ADN matrice est sous forme circulaire, il est amplifié par un processus de cercle roulant et il en résulte un très long homopolymère de la séquence amplifiée (Gill and Ghaemi, 2008). Le système n'utilise pas d'amorces spécifiques, similaires aux oligonucléotides utilisés en PCR, mais un mélange d'hexamères aléatoires (NNNNNN) susceptibles d'amorcer la synthèse en tout point d'un ADN simple brin déjà produit. Il en résulte une amplification « en buisson » qui conduit à un long concatémére ramifié. Le taux d'amplification peut atteindre 10.000 fois la quantité de l'ADN matrice (Fig. 1).

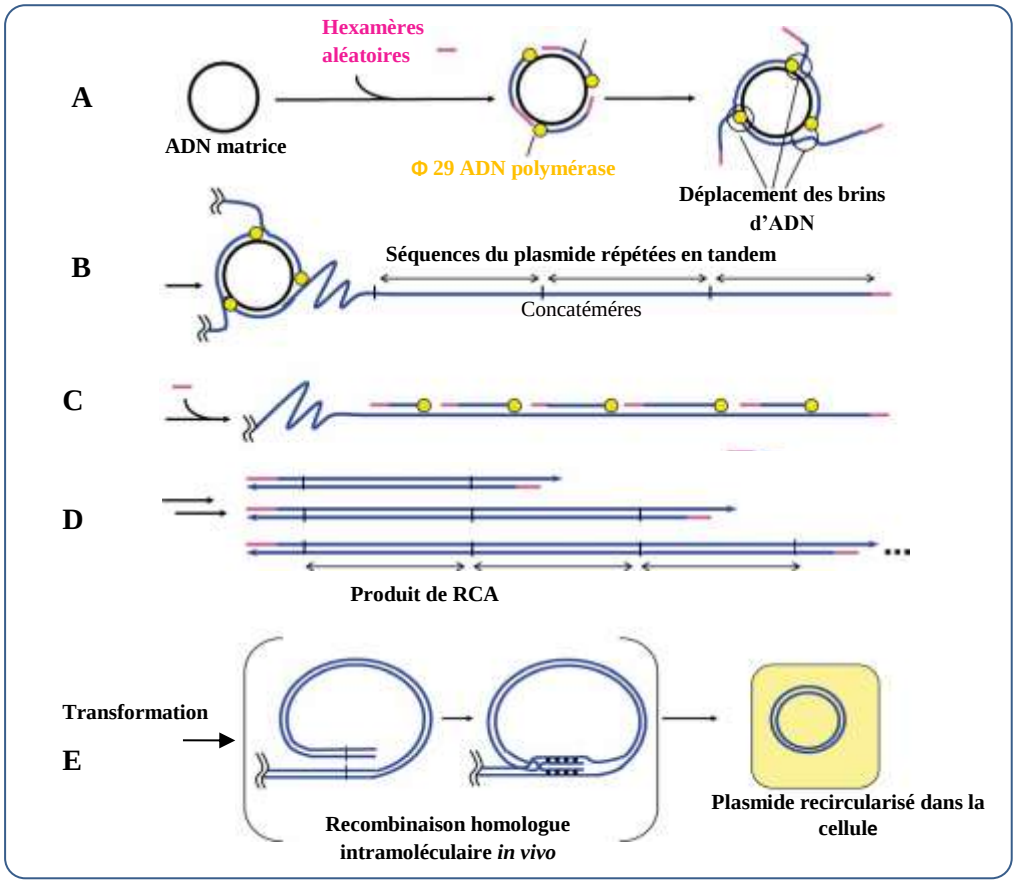


Fig. 1 : Mécanisme général d'une réaction de RCA : A : Formation du complexe ADN simple brin matrice polyhybridé avec la Φ29 polymérase. B : Le processus d'extension couvre plusieurs fois l'ADN matrice, grâce à la capacité de déplacement de la polymérase, produisant de longues molécules d'ADN de séquences répétées : « concatémères ». C : Les hexamères s'hybrident avec les concatémères pour les amplifier. D : Produit de l'amplification isotherme : Les concatémères de différentes longueurs sont par la suite transformés dans des bactéries ou des levures. E : Dans les cellules hôtes, les produits de la RCA sont recircularisés *in vivo* par recombinaison homologue.

L'amplification est non spécifique mais toute séquence longue ou circulaire est fortement privilégiée. Cette approche a été le plus souvent utilisée pour des diagnostics basés sur l'amplification sélective de cercles d'ADN réalisée par le rapprochement de deux oligonucléotides greffés sur des anticorps ou encore pour la détection de polymorphismes d'un seul nucléotide (single nucleotide polymorphisms : SNP) (Demidov, 2002).

Une autre variante de cette technique a été présentée dans le but d'introduire des mutations aléatoires lors de l'amplification d'ADN : *error-prone RCA* (Fujii, Kitaoka and Hayashi, 2006). Les mutations sont induites par ajout de $MnCl_2$ ayant pour effet de réduire la fidélité de la polymérase. Le produit de cette technique est une molécule d'ADN linéaire double brin composée de séquences répétées en tandem : des concatémères avec des mutations ponctuelles aléatoires. Ce produit est directement utilisé pour la transformation de bactéries ou levures où il est recircularisé *in vivo* par simple recombinaison homologue (Fig. 2).

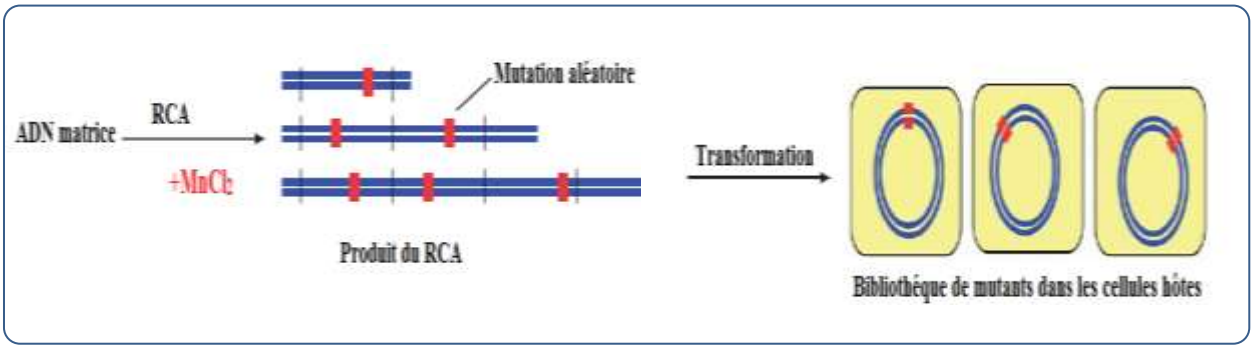
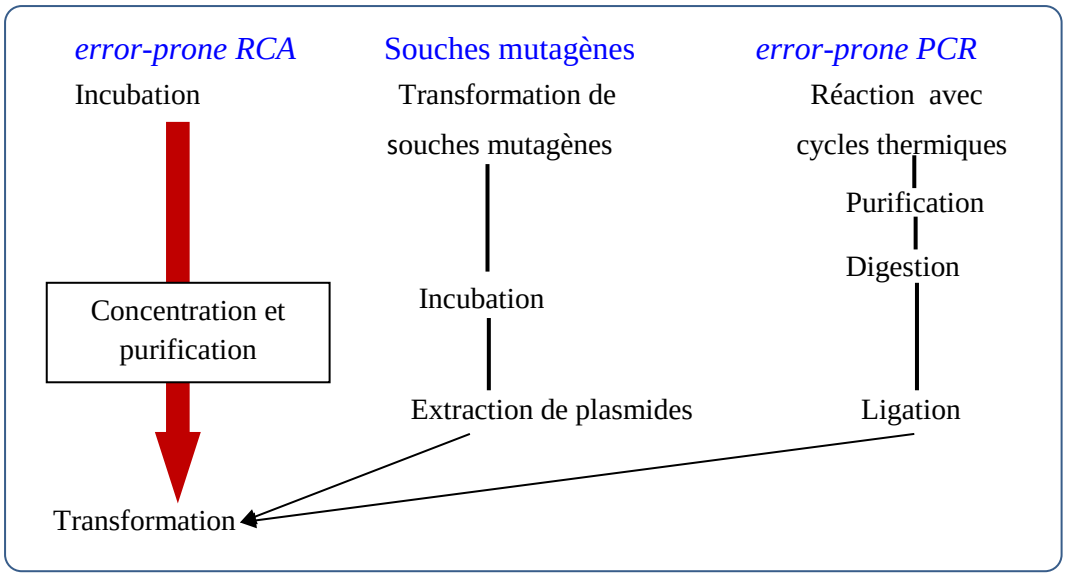


Fig. 2 : Schéma général d'*error-prone RCA* : L'ADN matrice est amplifié en présence de $MnCl_2$ réduisant ainsi la fidélité de la polymérase et induisant des mutations ponctuelles aléatoire. L'ADN produit et recircularisé *in vivo* permettant de construire une bibliothèque de mutants.

Cette technique présente plusieurs avantages : simplicité et rapidité de la mise en œuvre, pas de design d'amorces particulières et pas d'optimisation des cycles de température. Ci-dessous un diagramme comparatif entre RCA et d'autres techniques d'amplification (Diag. 1).



Diag. 1 : Comparaison des étapes d'amplification et mutagénèse dans RCA et dans des méthodes classiques de mutagénèse : Les étapes nécessaires pour l'amplification et l'introduction de mutations sont moins nombreuses en RCA par comparaison aux méthodes classiques de mutagénèse *in vitro* (*error-prone PCR*) et *in vivo* (les souches mutagènes).

d) Mutagenèse à saturation :

Quand la position d'un acide aminé est critique pour une fonction protéique quelconque, il est alors important d'identifier la nature de l'acide aminé idéal pour occuper cette position. Dans cette perspective plusieurs méthodes ont été développées. En effet, la mutagenèse à saturation (*Site Saturation mutagenesis* : SSM) permet de substituer ces sites précis par les 20 acides aminés (Labrou, 2010). La SSM peut être réalisée de deux manières différentes : soit en une seule étape de PCR soit en 2 étapes de PCR (la méthode des extrémités chevauchantes).

La première façon de procéder nécessite une seule étape de PCR avec 2 amorces contenant des codons mutés (Matsumura and Rowe, 2005). Le principe est celui utilisé classiquement pour la mutagénèse dirigée, par exemple dans le kit *Quick change* commercialisé par *Stratagene*, mais les oligonucléotides sont dégénérés à la position à modifier. Dans ce cas, les séquences des amorces sont en partie complémentaires aux brins opposés du plasmide et contiennent des mésappariements au niveau des codons à muter. Le plasmide est amplifié par PCR produisant ainsi des plasmides avec des mésappariements ou « *mismatch* » à la position désirée. Le plasmide sauvage ayant servi de matrice est éliminé par digestion avec une enzyme de restriction (DpnI). Cet enzyme nécessite un site de reconnaissance méthylé pour être actif ce qui est le cas de l'ADN matrice mais pas de l'ADN

amplifié. Les mésappariements des plasmides sont réparés, après transformation, *in vivo* par la machinerie bactérienne (Fig. 3).

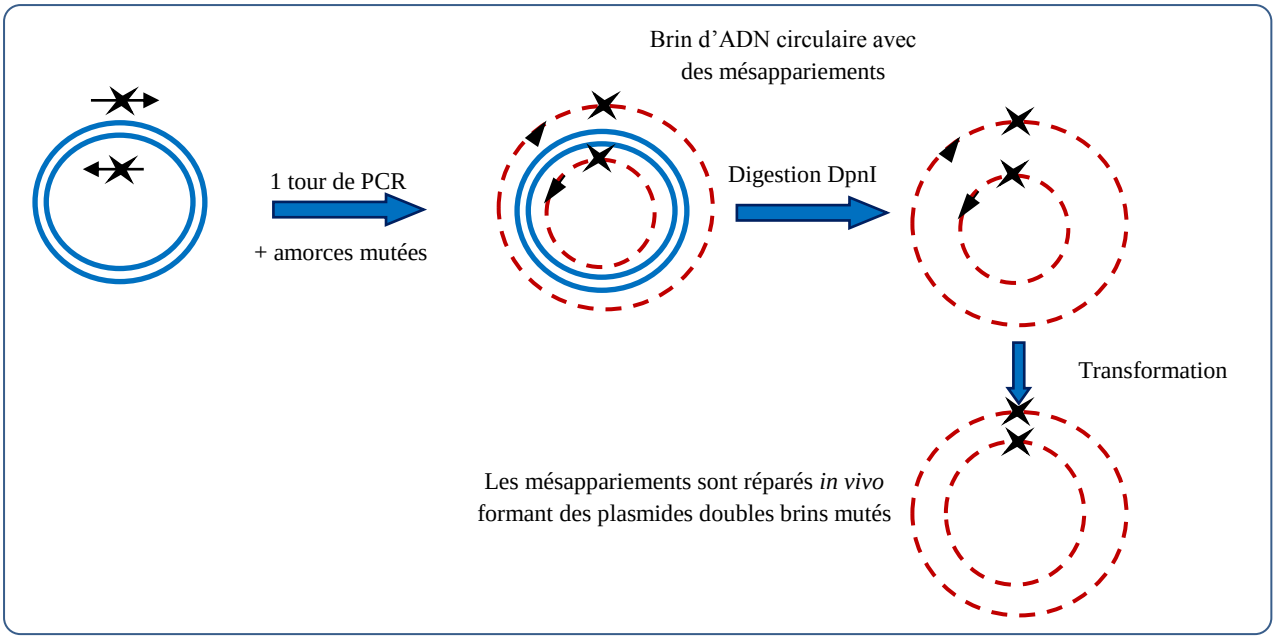


Fig. 3 : Mutagenèse à saturation en une seule étape de PCR.

La deuxième façon de procéder est la méthode des extrémités chevauchantes (*overlap extension method*) (Alcalde et *al.*, 2006). Elle nécessite deux amplifications du gène d'intérêt avec deux paires d'amorces. La première paire (2-3) contient les codons portant la mutation et les amorces vont s'hybrider formant un mésappariement (*mismatch*) alors que la deuxième paire (1-4) est entièrement complémentaire à la séquence cible. Les 2 PCR sont réalisées indépendamment et sont, par la suite, mélangées. Après dénaturation et réhybridation, une population d'ADN double-brins hétérogènes hybridés au niveau de la région mutée et comprenant des bouts cohésifs, est générée. Les bouts manquants sont, alors, complétés par simple polymérisation à l'aide de l'ADN polymérase. Le produit obtenu est amplifié par PCR avec les amorces (1-4) (Fig. 4). Cette technique peut être appliquée à un plasmide entier avec l'avantage d'une seule étape de PCR. Mais elle ne reste efficace que pour des plasmides de taille inférieure à 10kb.

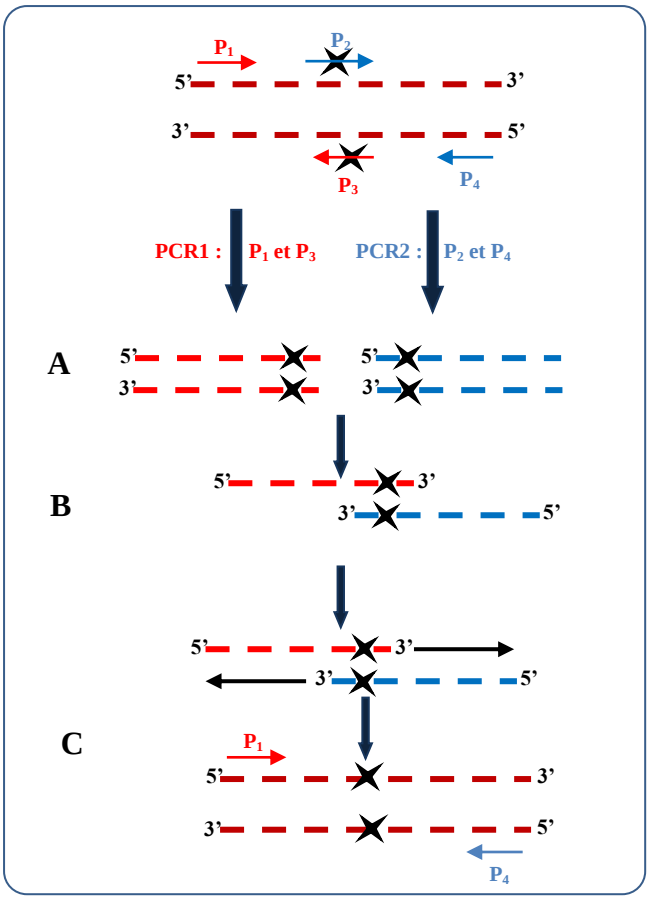


Fig. 4 : Schéma général de la méthode des extrémités chevauchantes : A : Les 2 PCR réalisées séparément avec les 2 paires d'amorces (1-3) d'une part et (2-4) d'autre part. B : Dénaturation et réhybridation des deux produits de PCR produisent une population de molécules d'ADN hétérogènes qui sont par la suite complétées par l'ADN polymérase. C : La dernière étape d'amplification par les amorces (1- 4) produit des séquences de départ avec les mutations codées par les amorces 2 et 3.

Plusieurs autres techniques ont été mises au point telles que la mutagénèse par des cassettes partiellement double-brins, comportant des codons aléatoires ou « randomisés », qui seront par la suite introduites dans le gène d'intérêt par simple ligation. Lors de ces mutagenèses, les codons aléatoires sont obtenus à partir d'oligonucléotides dégénérés. Les plus utilisés sont NNN, NNB, NNK ou NNS (où N = A/C/G/T, B = C/G/T, K = G/T et S = G/C). Cette combinaison permet de coder les 20 acides aminés. Cette approche peut conduire à l'introduction de codons stop dans la séquence, générant des protéines tronquées. En biaisant les pourcentages des codons, il est possible de limiter la présence de ces mutations (Labrou, 2010).

4.2. Méthodes de criblage et de sélection

La deuxième étape du processus d'évolution est l'étape de sélection des variants ayant incorporé la propriété d'intérêt. De la même manière, selon le mode opératoire, il existe des

méthodes de sélection *in vitro* et d'autre *in vivo*. Ces méthodes sont essentiellement basées sur les propriétés d'interaction protéine-ligand (Amstutz et *al.*, 2001).

a) Méthodes de sélection in vivo

Les méthodes *in vivo* nécessitent une étape préalable de clonage des gènes mutés dans des plasmides ou des phagemides appropriés pour la transformation cellulaire (bactéries, levures). En effet si la sélection est basée sur une interaction intracellulaire, les techniques utilisables reposent sur les approches de doubles hybrides chez la levure et les approches de complémentation de fragments protéique ou PCA (*Protein-fragment Complementation Assay*). Dans le cas où la sélection est basée sur une simple interaction protéine-ligand les techniques développées sont l'exposition sur cellules (bactérie ou levure), l'exposition sur phages et sur d'autres types de virus.

- *Technique du double hybride :*

C'est une technique développée par Fields et Song en 1989. C'est un système génétique visant à étudier des interactions protéines- protéines se basant sur les propriétés de la protéine GAL4 de la levure *Saccharomyces cerevisiae*. Cette protéine est un activateur de la transcription, essentielle pour l'expression des gènes codant les enzymes de dégradation du galactose. Elle se compose de deux domaines fonctionnels séparables : un domaine N-terminal qui se lie spécifiquement à la séquence d'ADN UASG, et un domaine C-terminal contenant des résidus nécessaires pour activer la transcription. Le système généré comporte deux protéines hybrides. La première contient le domaine d'interaction avec l'ADN fusionné à une protéine X. La deuxième comprend le domaine d'activation de la traduction de la GAL4 fusionné à une protéine 'Y'. Par la suite, si X et Y peuvent former un complexe et reconstituer ainsi la protéine GAL4, il y aura induction de la transcription d'un gène rapporteur qui sera régulé par la séquence UASG. Le système peut être appliqué à une grande variété de protéines, plus particulièrement pour identifier une protéine interagissant avec une protéine connue par simple sélection sur milieu galactose (Fields and Song, 1989).

- *PCA « Protein Fragment complementary assay »*

La PCA est une technique développée par l'équipe de S. Michnick, dans le but d'établir une approche expérimentale permettant de répondre à la série de questions suivantes : comment, quand, où et dans quelles conditions les protéines sont

activées/désactivées ou encore interagissent ? (Michnick, 2003) Le principe de la PCA est le suivant : 2 protéines à tester sont fusionnées chacune à un fragment d'une protéine rapporteuse et sont exprimées dans une même cellule. Si ces deux protéines interagissent la protéine rapporteuse est alors reconstituée rétablissant ainsi sa fonction naturelle (Michnick, 2001) et permettant la détection des cellules ayant les deux partenaires d'interaction. La protéine rapporteuse peut être, soit une enzyme soit une protéine fluorescente. En effet, plusieurs protéines ont été utilisées pour cette application : on peut citer la dihydrofolate réductase (DHFR), l'aminoglycoside kinase, la TEM β-lactamase et la protéine à fluorescence verte (*Green Fluorescent Protein : GFP*). La particularité de ces protéines est que leurs fragments sont incapables d'interagir spontanément et reformer la protéine entière (Fig. 5), le but est qu'ils le deviennent par la reconnaissance des partenaires auxquels ils sont fusionnés.

Cette technique permet de détecter directement des interactions protéines – protéines. Elle a l'avantage d'exprimer les protéines sous leur forme native, avec les modifications post-traductionnelles si le système d'expression le permet, et de les localiser dans les cellules. Par ailleurs, cette approche peut être réalisée dans tous types cellulaires, dans les bactéries comme dans les cellules eucaryotes. Dans les bactéries, cette technique permet de créer des librairies plus diverses que celles obtenues par double hybride chez *S. Cerevisiae*.

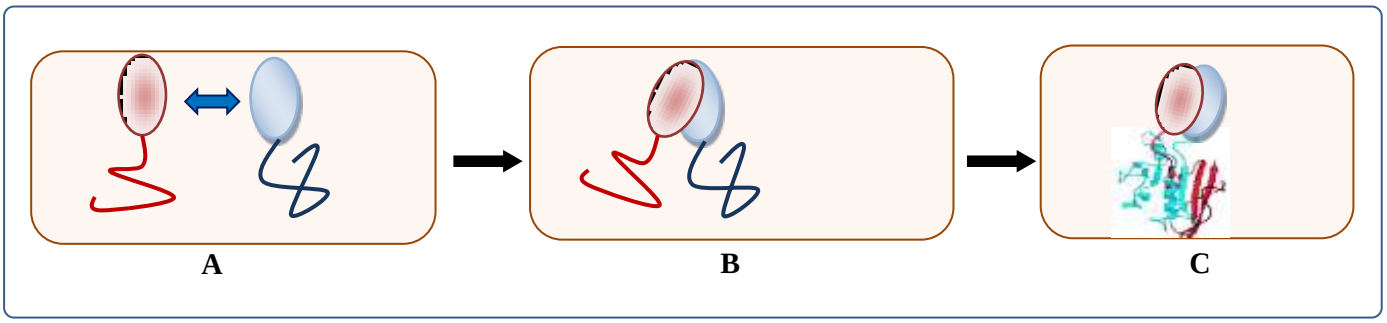


Fig. 5 : Schéma général de la PCA : A : Les 2 protéines d'intérêt sont exprimées dans la même cellule, fusionnées chacune à un domaine de la protéine rapporteuse (enzyme ayant une activité vitale pour la cellule ou un effet fluorescent). B : Si les protéines d'intérêt interagissent les 2 domaines de la protéine porteuse s'approchent. C : la protéine porteuse (ici DHFR) retrouve sa conformation naturelle à partir des 2 domaines et sa fonction est rétablie.

- *Exposition sur cellules : Cell Display*

Grâce au développement des systèmes d'expression, des polypeptides sont efficacement exposés à la surface des cellules. Selon la nature des protéines à faire évoluer et du système d'expression mis en œuvre, les cellules support utilisées peuvent être des bactéries, des levures ou encore des cellules de mammifères (Fig. 6).

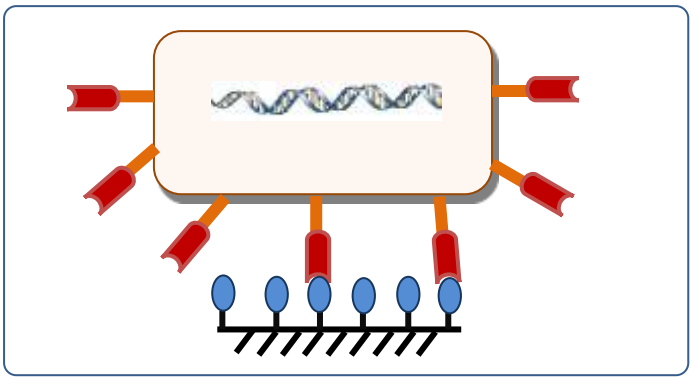


Fig. 6 : Exposition sur cellule.

A titre d'exemple, ce système a permis de cribler, une bibliothèque de variants de protéase de membrane externe OmpT, exposée à la surface d'*E.coli* et d'isoler des enzymes qui sont plus actives que les clones de départ. En effet, l'équipe est partie d'une bibliothèque de $6*10^5$ clones exposant à leur surface l'enzyme OmpT préalablement mutée et, par tri de cellules par fluorescence (FACS) et un substrat peptidique fluorescent, il a été possible de sélectionner des variants 60 fois plus actifs. Les bactéries sont isolées grâce au clivage du substrat synthétique avec lequel elles sont incubées. Le substrat contient un colorant fluorescent, une séquence de clivage (Arg-Val) et un partenaire qui éteint le signal de fluorescence avant clivage (Olsen et *al.*, 2000).

Les levures ont été utilisées en tant que support, la première fois par l'équipe de Witthrup, pour l'exposition d'une bibliothèque relativement petite ($3*10^5$clones) de scFV aléatoirement mutés. Par un système de criblage par cytométrie de flux, des mutants de scFV spécifiques à la région Vβ8 du récepteur des cellules T ont été isolés (Kieke et *al.*, 1997). Ce système permet d'obtenir directement des variants isolés sans nécessiter d'étapes supplémentaires de sous-clonage et de surexpression. L'exposition à la surface de la levure a aussi l'avantage de pouvoir cribler des protéines eucaryotes de haut poids moléculaire, qui peuvent être glycosilées ou présenter des ponts disulfures, ce qui est impossible à obtenir avec les bactéries. Bien que plusieurs cellules hôtes soient disponibles, *Saccharomyces cerevisiae* reste l'organisme le plus communément utilisé (Shibasaki and Ueda, 2010). Une grande variété de protéines procaryotes et eucaryotes ont été exposées à la surface des levures : des enzymes ont été exposées pour des applications en bioconversion, mais aussi des anticorps et des récepteurs pour des applications analytiques et en bioséparation. A titre d'exemple, l'exposition et la sécrétion chez la levure a permis l'amélioration de la production de bioéthanol (Kondo and Ueda, 2004).

Grâce à ces avancées, Ho et Pastan ont décrit une stratégie de maturation d'affinité d'anticorps et de leurs dérivés par exposition sur cellules de mammifères. Deux protéines, la mésothéline et CD22, surexprimées dans les cellules tumorales, ont été utilisées comme cibles lors de la sélection d'anticorps. Moyennant des systèmes de criblage et de sélection sur des cellules rénales embryonnaires humaines 293T (*Human Embryonic Kidney 293T* : HEK-293T), des scFV et des IgG entiers liant spécifiquement la mésothéline et le CD22 avec une haute affinité ont été isolés. Les affinités déterminées par cytométrie de flux sont de 2.5 nM contre 5.8 nM pour le type sauvage (Ho and Pastan, 2009).

- *Exposition sur phage : Phage Display*

Cette technique permet l'expression de peptides ou de protéines exogènes à la surface des particules de phages, dans le but de sélectionner et d'amplifier un polypeptide capable d'interagir avec une molécule cible choisie. Cette technique est probablement la plus utilisée dans la découverte de nouvelles interactions spécifiques à partir de répertoires de peptides ou polypeptides.

Le concept a été décrit pour la première fois par Smith en 1985 (G. P. Smith, 1985) lorsqu'il est apparu que le gène III du virion du phage codait pour une protéine mineure de l'enveloppe virale. Cette protéine est constituée de deux domaines : un domaine N-terminal qui se lie au pilus F lors de l'infection de la bactérie et un domaine C-terminal enfoui dans le virion participant à sa morphogenèse. L'observation initiale était que l'insertion d'un ADN étranger entre les 2 domaines de la protéine III semblait ne pas perturber le fonctionnement de la protéine et que la protéine synthétisée était exposée et accessible à la surface du phage. Cette approche a d'abord été décrite comme une approche d'identification de gènes naturels clonés dans des vecteurs dérivés du phage M13.

L'innovation majeure a été de construire, sur le même principe, des bibliothèques de peptides de séquences aléatoires et de montrer qu'il est alors possible de « cloner » des séquences artificielles susceptibles d'interagir efficacement avec des protéines cibles choisies. L'approche permettait, d'une façon très générale, la sélection et l'amplification de séquences, même très rares, au sein de bibliothèques de grande diversité. Plusieurs améliorations ont été apportées ultérieurement au système décrit par Smith en le rendant généralisable (Fig. 7). Le développement de cette méthodologie a joué un rôle déterminant dans le développement des stratégies combinatoires en ingénierie des protéines.

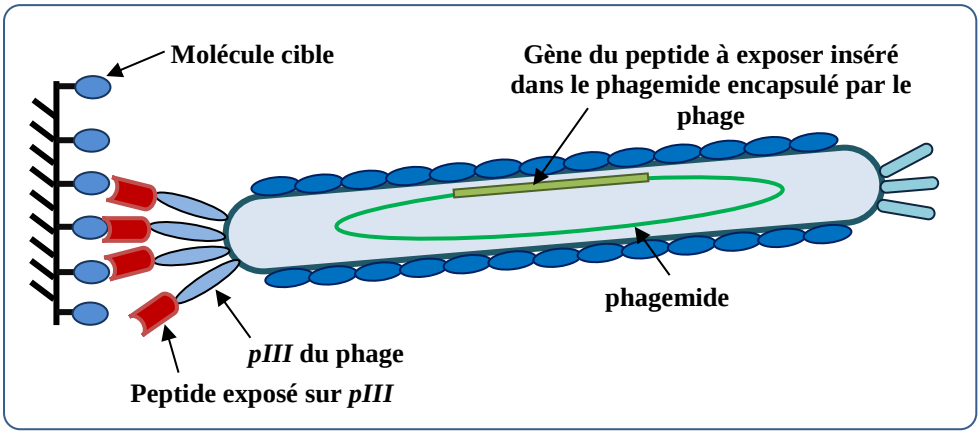


Fig. 7 : Exposition de peptides sur le phage M13.

Pourquoi les phages et quel type de phages ?

Les phages utilisés pour l'exposition des protéines et des peptides sont des phages filamenteux spécifiques des entérobactéries de type M13, fd ou f1, choisis notamment pour la simplicité de leur organisation génétique et structurale ainsi que pour la facilité de leur production. Parmi les caractéristiques intéressantes pour leur utilisation dans les méthodes de *phage display* (Pande, Szewczyk, and Grover, 2010), on peut citer :

- un génome de taille relativement modeste (quelques kb) qui peut être purifié très facilement soit sous forme double brin à partir de cellules infectées soit sous forme simple brin à partir des particules virales.
- la taille de l'ADN encapsidé peut être de taille variable et donc l'encapsidation n'est pas perturbée par insertions de séquences exogènes dans le génome viral.
- les protéines de l'enveloppe sont modifiables et peuvent être exprimées, dans certaines conditions, sous forme de protéines de fusion sans affecter l'infectiosité des phages.
- les particules virales sont très résistantes et tolèrent une large gamme de pH, de température ou de dénaturant, d'où leur capacité à s'adapter à différentes conditions de sélection.
- les phages filamenteux ne lysent pas les cellules infectées qui continuent de croître après infection. Les phages sont présents à une densité élevée dans le surnageant d'une culture de cellules infectées (environ 10^{11} cfu/mL)

Le phage le plus communément utilisé est le phage M13. Et les protéines d'intérêt peuvent être fusionnées à l'extrémité N-terminale des protéines pIII, pVII, pVIII, pIX, à

l'extrémité C-terminale des protéines pVI, pIII et pVIII (Sidhu, 2001) ou encore à des formes modifiées de pVIII. La pIII et la pVIII restent de loin les plus utilisées dans les processus de sélection. La pIII est une protéine de 406 acides aminés, présente à l'état de quelques copies à une extrémité de la particule virale, tandis que la pVIII est plus petite (50 acides aminés). Cette dernière s'assemble pour former le tube creux au sein duquel est encapsidé l'ADN, et se trouve donc répétée de très nombreuses fois tout le long du filament du virion.

Phages et phagemides ?

Initialement, il semblait que seuls des peptides exogènes de taille réduite (7 ou 11 résidus) pouvaient être fusionnés avec les protéines virales. Une protéine fusionnée de taille trop importante risquait d'interférer par encombrement stérique avec l'assemblage du virus en étant exposée sur toutes les protéines de l'enveloppe. Elle peut aussi interférer avec les capacités infectieuses du phage comme étant exprimée à l'extrémité impliquée dans l'adhésion aux bactéries. Ce problème a pu être contourné en produisant des virions hybrides dont une partie seulement des protéines virales sont fusionnées avec les protéines d'intérêt, tandis qu'une autre partie, provenant d'un gène non modifié, permet l'assemblage ou l'infectiosité du phage. Apporter une copie intacte et une copie fusionnée est réalisable expérimentalement de deux manières : soit en apportant les protéines de l'enveloppe sous deux formes (une sauvage et une en protéine de fusion) dans le génome du virus soit en intégrant le gène de la protéine de fusion dans un phagemide (Soltes et *al.*, 2003).

Le phagemide est un plasmide qui comporte d'une part la séquence nécessaire à la réplication de l'ADN viral et à son encapsidation au sein des virions, et d'autre part le gène des protéines à exposer, sous forme fusionnée avec le gène de la protéine virale utilisée (pIII ou pVIII). Les autres gènes, indispensables à la réplication et à l'assemblage du phage, sont apportés par un phage auxiliaire ou « *helper phage* ». Ce dernier est un phage qui porte le gène sauvage de la pIII ou la pVIII nécessaire puisque l'autre copie portée par le phagemide ne suffit pas à la viabilité des phages (Rondot et *al.*, 2001). Une bactérie, co-infectée par le phagemide et le *phage helper* va produire des particules virales sauvages et d'autres hybrides exposant des protéines exogènes. Ces particules encapsulent le phagemide préférentiellement au génome du phage auxiliaire (Fig. 8). Il y a donc une sorte de compétition entre la protéine sauvage et la protéine de fusion.

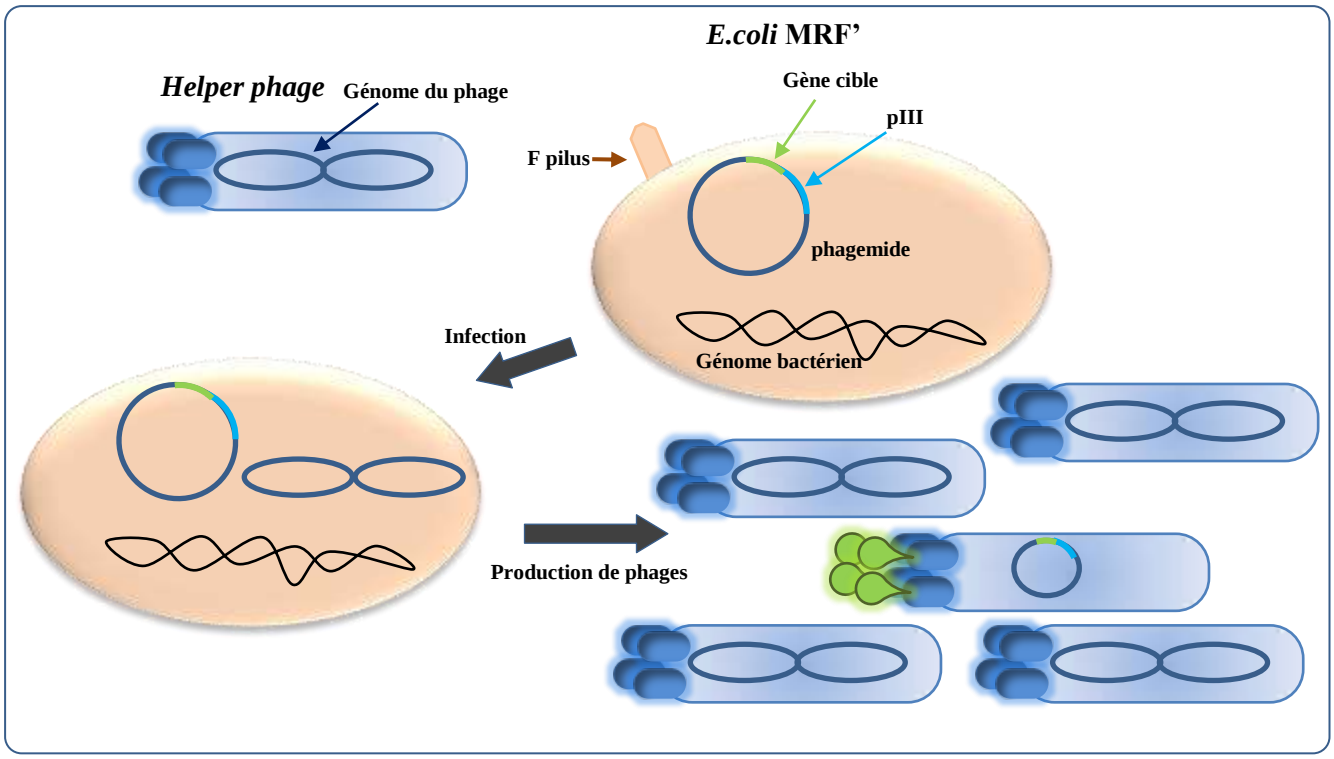


Fig. 8 : Principe général de production de phages exposant la protéine d'intérêt.

Les principes généraux d'exposition des protéines, décrits ci-dessus, par le biais du système de type 3 (selon la classification de Smith (George P. Smith and Petrenko, 1997)) sont vrais pour l'exposition des protéines sur phages par la protéine pVIII. La seule différence à noter pour le système de type 8 est la valence d'exposition. En effet, si la pIII fusionnée permet d'exposer au plus 5 copies par virion la pVIII quant à elle permet l'exposition de centaines voire de milliers d'exemplaires de protéines de fusion par particule de phage (Bratkovič, 2009). L'utilisation d'un système ou d'un autre dépend essentiellement des protéines à exposer sur les phages. En effet, un interacteur peut être sélectionné par le système de type 8 même avec une faible affinité puisque la sélection est aidée par l'effet d'avidité résultant d'une présentation multicopies. Ce système a parfois été utilisé avec succès et l'avidité n'a pas empêché de sélectionner des anticorps anti-carbohydrate de bonne affinité avec le système de type 8 (k_D = 10nM) (Dinh et *al.*, 1996).

Le taux d'exposition ou le taux de « *display* » est aussi influencé par les conditions de cultures. En effet, avec un phagemide de type 3, la majorité des phages n'exposent pas de protéines de fusion (Kramer et *al.*, 2003). La protéine III intacte provenant du phage auxiliaire est produite et/ou assemblée avec les particules virales beaucoup plus efficacement que les protéines III fusionnées de sorte que la majorité des phages n'exposent que la protéine III sauvage. Seule une petite fraction (souvent de l'ordre de 0,1 à 1% selon les conditions

utilisées) des particules virales comporte une copie de la protéine fusion (et donc seule cette fraction est sélectionnable). Une fraction très faible expose deux copies ou plus, de sorte que l'exposition de la protéine est alors essentiellement monovalente, ne créant pas l'avidité des protéines exposées. Plusieurs paramètres peuvent influencer le taux d'exposition : la séquence d'export vers l'espace périplasmique des bactéries, la nature et le patrimoine génétique du « *phage helper* » et le type de promoteurs utilisés pour l'expression (Rondot et *al.*, 2001).

Il est vrai que le phage M13 est le plus communément utilisé, en particulier pour les protéines qui peuvent être exportées dans l'espace périplasmique bactérien. Elles peuvent se replier correctement (environnement oxydant). Mais il existe aussi d'autres phages qui ont servi à l'exposition de protéines. La protéine D (pD) du phage *Lambda*, par exemple, a été fusionnée, aussi bien à son extrémité N que C-terminale, à des polypeptides et des protéines multimériques qui se replient efficacement dans le milieu réducteur du cytoplasme. Ces phages ont aussi servi pour exposer des bibliothèques d'ADNc pour diagnostiquer des tumeurs (Minenkova et *al.*, 2003). Un autre phage lytique, le phage T7, a été aussi utilisé pour exposer des protéines et d'autres molécules pour des applications entre autres thérapeutique. La protéine majeure de la capside de ce phage est la protéine 10 qui existe sous deux formes : 10A et 10B (Bratkovič, 2009). Rosenberg et ses collaborateurs ont proposé un système d'exposition de protéines en fusion avec la protéine 10B tronquée à son extrémité C-terminale. Cette technologie a été par la suite commercialisée par *Novagen*. Le phage T4 a été aussi utilisé pour exposer un IgG actif à la surface de sa capside (Ren and Black, 1998).

- *Exposition sur d'autres types de virus*

D'autres types de virus ont été adaptés pour l'exposition de peptides et même de protéines et pour l'établissement de stratégie de sélection et de criblage de banques. Un exemple d'utilisation de virus eucaryote a été proposé par l'équipe de Kan (Kasahara, Dozy and Kan, 1994). Cette équipe a exploité une approche d'infection de cellules hôtes par un virus induite par une interaction antigène-anticorps. Pour cela, l'érythropoïétine (EPO) est exprimée au niveau de l'enveloppe protéique du virus Moloney de la leucémie murine (Mo-MuLV). Cela a été obtenu en substituant une région de la séquence N-terminale du gène de l'enveloppe par le gène de l'EPO, tout en conservant la portion du gène codant pour le peptide signal des protéines de l'enveloppe du côté N-terminal et les résidus Cystéine C-terminaux indispensables à l'encrage à l'enveloppe interne du virus. Ainsi, l'infection virale est améliorée non seulement des cellules murines mais aussi des cellules érythrocytaires

humaines qui exposent un récepteur à l'EPO. Le vecteur rétroviral construit et l'approche peuvent être généralisés et appliqués à une large gamme de virus qui exposent un ligand quelconque spécifique d'un récepteur. Cette stratégie donne la possibilité de délivrer les gènes dans les tissus ou les organes afin de permettre le traitement de maladie génétiques et du cancer.

Bien que ces approches soient efficaces et très utilisées, leur diversité réelle et leur richesse potentielle est liée à leur mode de construction. En effet, l'étape de transformation inévitable limite la taille des librairies construites. Cette dernière est au maximum de l'ordre de 10^{10} clones indépendants pour les librairies transformées dans *E.coli* et une taille plus réduite lors que la transformation est réalisée dans la levure. De plus, une sorte de « présélection » est imposée par l'environnement des cellules hôtes. En effets le nombre de variants d'une banque peut être réduit par une éventuelle toxicité des protéines exprimées qui entraine un ralentissement de la croissance cellulaire ou peut même avoir un effet létal. Ceci peut biaiser le processus de sélection : un clone qui croit plus vite est un clone qui est plus représenté dans la banque. Ces inconvénients peuvent être contournés par les méthodes de sélection *in vitro*.

b) Méthodes de sélection in vitro

Bien que les méthodes de sélections *in vitro*, décrites ci-dessous, soient des méthodes de sélection et de criblage de banques de protéines, il est important de mentionner qu'elles sont inspirées par une méthode de sélection d'acides nucléiques. C'est la technique proposée en 1990 (Ellington and Szostak, 1990) intitulée évolution systématique de ligand par enrichissement exponentiel (Systematic Evolution of Ligands by EXponential Enrichement : SELEX). La technique SELEX consiste à réaliser plusieurs cycles comportant une étape de transcription d'un pool d'acide nucléique aléatoires, suivie d'une étape de sélection par contact avec la molécule cible et finalement une étape de RT-PCR permettant une amplification exponentielle des acides nucléiques sélectionnés. Après les cycles de sélection, les interacteurs obtenus sont clonés, séquencés et caractérisés. Cette technique présente l'avantage d'avoir l'information génotypique et phénotypique sur la même molécule et d'obtenir rapidement des interacteurs hautement spécifiques à partir de très larges librairies initiales (10^{15} à 10^{16} séquences). Un des inconvénients de cette méthode est la sensibilité des

ARN aux RNases ubiquitaires mais l'inconvénient majeur réside dans le fait que l'ARN est une molécule polyanionique interagissant préférentiellement avec des cibles positives ce qui restreint la gamme de cibles que l'on peut choisir. Pour contourner ce problème et élargir le répertoire des interactions chimiques, les ARN ont été remplacés par les peptides et les protéines d'où les méthodes de sélection protéiques *in vitro* (Stoltenburg, Reinemann, and Strehlitz, 2007).

- *Exposition sur ribosome : Ribosome Display*

Juste après la publication de la technique SELEX, un brevet a été déposé proposant une approche similaire appelée « *Ribosome Display* » permettant l'enrichissement de peptides à partir d'une bibliothèque. Mattheakis et ses collaborateurs ont utilisé, pour la première fois, cette méthode expérimentale (Mattheakis, Bhatt, and Dower, 1994) pour identifier parmi des séquences aléatoires codant des décapeptides, celles qui reconnaissent un anticorps monoclonal spécifique de la dynorphine B en utilisant le système de transcription S30 d'*E.coli*. Par la suite l'approche a été optimisée dans le but de l'adapter à l'exposition de protéines complètes correctement repliées. Ces travaux sur l'évolution et la sélection d'un fragment monochaîne d'anticorps (scFv) ont été réalisés par Hanes et Plückthun (Hanes and Plückthun, 1997).

L'identification, par *Ribosome Display*, d'interacteurs spécifiques pour une molécule cible se fait en plusieurs cycles de sélection. Un cycle de sélection comporte différentes étapes dont la première est la formation d'un complexe ribosomal (ARNm + Ribosome + chaîne polypeptidique) (Fig. 9). La première étape de formation du complexe ribosomique peut se faire de deux façons différentes : soit en couplant transcription-traduction partant d'une bibliothèque d'ADN, soit en préparant l'ARNm par transcription et purification préalables pour passer par la suite à l'étape de traduction et formation du complexe. A ce stade, il y a formation de complexes ribosomaux qui contiennent une protéine repliée exposée mais qui reste liée par sa séquence C-terminal à l'ARNt, qui est lui-même encore complexé au ribosome. La traduction est par la suite arrêtée par refroidissement de la réaction en ajoutant du tampon froid contenant du Mg^{2+}. Ce traitement favorise la condensation des ribosomes empêchant ainsi la dissociation ou l'hydrolyse des complexes. Une fois la bibliothèque exposée sur les ribosomes, le mélange est incubé en présence de la cible. Cette dernière peut être immobilisée directement sur la surface d'un tube, d'une plaque ou encore biotinylée et fixée à la surface de billes magnétiques couvertes de streptavidine. L'avantage de cette

pratique est de s'assurer que la molécule cible biotinylée est efficacement capturée dans sa conformation naturelle et non déformée par d'éventuelles interactions avec le support. Par ailleurs, il est ainsi possible de contrôler la concentration exacte de cible utilisée pour la sélection. Les complexes peu ou pas spécifiques sont éliminés par des lavages avec du tampon contenant du magnésium. Par la suite, l'élution peut être réalisée soit par ajout d'un excès d'EDTA, ce qui élimine le Mg^{2+} dissociant ainsi les complexes. Les ARNm libres sont récupérés sans avoir recours à la dissociation des interacteurs protéiques. L'élution peut aussi être réalisée par compétition avec la cible libre ce qui permet d'isoler les ARNm correspondant aux interacteurs fonctionnels et spécifiques seulement. Enfin, Les ARNm isolés sont amplifiés par RT-PCR et l'ADN résultant est utilisé pour le cycle suivant. Une partie de cet ADN est analysée par clonage, séquençage et tests ELISA.

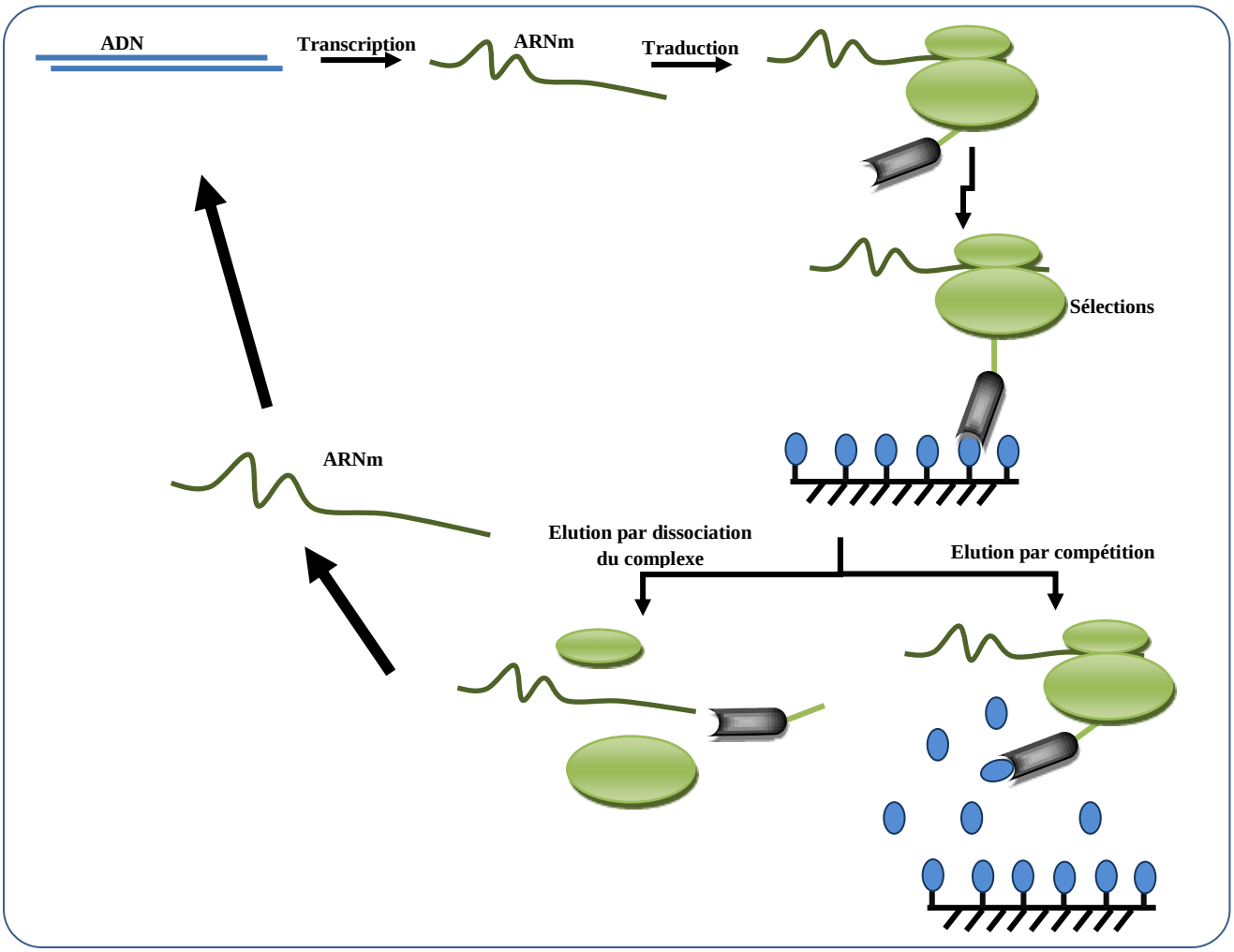


Fig. 9 : Principe général de la technique du *Ribosome Display*.

L'exposition sur ribosomes est une méthode ou chaque cycle peut être simultanément un cycle de sélection et un cycle d'évolution puisque l'étape de RT-PCR peut être mise à

profit pour réintroduire de la diversité au sein des populations qui sont sélectionnées. Le premier succès du *ribosome display* est la sélection à partir d'une bibliothèque de 10^{12} peptides (de 10 acides aminés), de variants pouvant reconnaître spécifiquement un anticorps monoclonal fixant la dynorphine B. En effet, après 5 tours de sélection par *ribosome display*, en utilisant un système de transcription-traduction couplées d'*E.coli*, plusieurs peptides fixant l'anticorps à des affinités de 7.2 à 140 nM ont été isolés.

Le système d'exposition sur ribosome d'*E.coli* a été par la suite optimisé et amélioré pour permettre un affichage plus efficace de protéines entières correctement repliées. En effet, Knappik et ses collaborateurs (Knappik et *al.*, 2000) ont pu isoler, à partir d'une large bibliothèque de fragments d'anticorps synthétiques de $2*10^9$ clones indépendants, différentes familles de scFV qui interagissent avec l'insuline. Par comparaison des séquences des fragments d'anticorps de la librairie de départ avec ceux obtenus après les 5 tours de sélection, l'équipe a constaté qu'aucun scFV de départ n'était retrouvé après la sélection. En effet, des mutations avaient été introduites au cours de l'étape d'amplification des ARN en particulier avec l'utilisation d'une polymérase sans activité correctrice. Ainsi, une nouvelle diversité a été générée à partir de laquelle une deuxième génération de mutants avait émergé. Ceci mime le processus d'hypermutation somatique naturelle des anticorps, observé lors de réponses secondaires après exposition multiple à l'antigène. La caractérisation biophysique de ces scFv a montré que leurs affinités pour l'antigène ont été ainsi augmentées 40 fois et peuvent atteindre des k_D de l'ordre du picomolaire pour certains.

D'autres travaux ont utilisé un système d'exposition sur ribosomes eucaryotes obtenus par exemple à partir de réticulocytes de lapin (He et *al.*, 1999). Ce système a permis de sélectionner des fragments d'anticorps humains fixant la progestérone à partir d'une bibliothèque préparée à partir de souris transgéniques. En effet, une première bibliothèque de fragments d'anticorps complétement humains a été générée par immunisation de souris transgéniques. Ces dernières sont porteuses de gènes humains de chaînes lourdes et légères alors que les fragments endogènes sont rendus silencieux. Une deuxième bibliothèque a été générée par combinaison aléatoire des chaînes H et L par PCR. Finalement les fragments spécifiques à la progestérone ont été sélectionnés par 5 tours de *ribosome display*. Les ADN correspondants ont été par la suite clonés et exprimés dans *E.coli*. De plus, l'étape d'amplification a été réalisée à l'aide de polymérase dotée d'une activité correctrice permettant d'avoir une méthode exclusivement de sélection.

- *La fusion Peptide –ARN : RNA-Peptide fusion ou In Vitro virus*

Cette approche proposée par Roberts et Szostak, décrit une méthode permettant de lier de manière covalente un peptide à l'ARN qui le code (Roberts and Szostak, 1997). En effet, l'ARNm est transcrit *in vitro* puis purifié et par la suite lié par son extrémité 3', à une séquence ADN *linker* marquée à la puromycine. Cette construction ADN-ARN-puromycine est purifiée puis traduite *in vitro*. Une fois la région ADN atteinte, les ribosomes se détachent du complexe permettant ainsi à la puromycine d'atteindre le site de la peptidyltransférase et de se lier covalemment au peptide. Les complexes hybrides obtenus sont utilisés dans l'étape suivante de sélection, ceux qui reconnaissent la cible sont élués et le matériel génétique est amplifié par RT-PCR. La seule étape critique de ce processus est l'étape de fixation de l'ADN marqué à l'ARNm mais son avantage majeur est la stabilité du complexe ce qui permet des conditions de sélection contrôlées et contraignantes (Fig. 10).

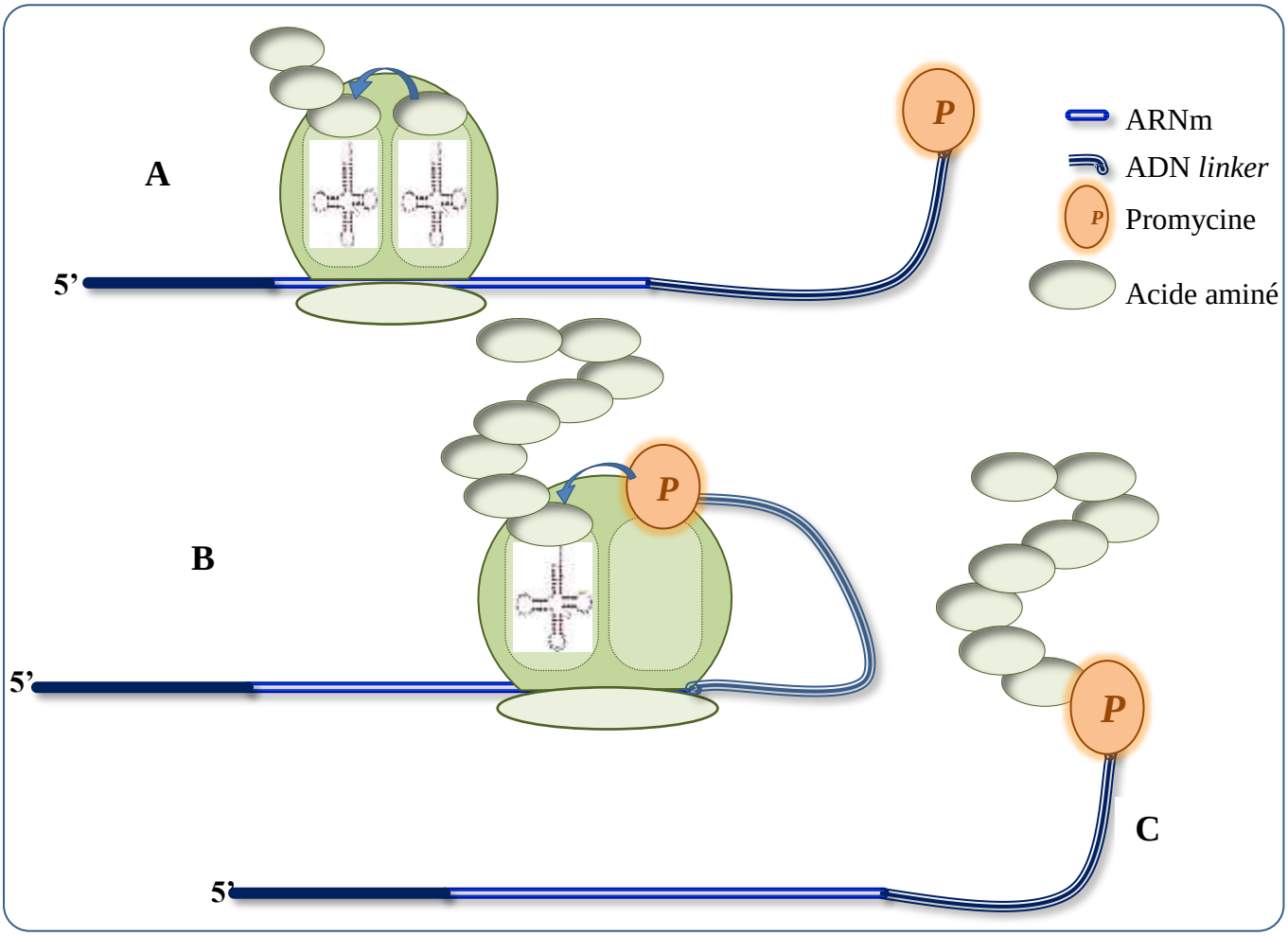


Fig. 10 : Schéma général de la formation du complexe ARNm-peptide : A : Initiation de la traduction et évolution des ribosomes au long du ARNm. B : Au niveau de la jonction ADN*linker*-ARNm, la puromycine est intégrée au niveau du site A du ribosome où elle est liée à la chaîne polypeptidique. C : Complexe ARNm-peptide obtenu après chromatographie d'affinité.

Des travaux récents se sont intéressés à la sélection de peptides de 16 acides aminés, capables d'interagir avec la Bcl-X(L) grâce à l'exposition sur l'ARNm (N. Matsumura et *al.*, 2010). Bcl-X(L) est un anti-apoptotique de la famille des Bcl-2 : une protéine mitochondriale inhibant d'activation de Bax et Bak. Cette équipe a même réussi à isoler des peptides plus affins que la protéine naturelle (IC50 = 0.9 µM pour le peptide dégénéré contre 11.8 µM pour le domaine naturel BH3 de Bak connu pour sa capacité à reconnaitre Bcl-X_L). Par la suite une protéine hybride a été construite pour laquelle le domaine BH3 de Back a été remplacé par le peptide identifié au sein de la bibliothèque. Cette protéine a gardé la capacité de fixer spécifiquement Bcl-X_L, d'inhiber son activité et même de masquer son effet dans les cellules saines.

Cette technique présente l'inconvénient de la fragilité des molécules d'ARNm par rapport aux ribonucléases particulièrement quand les cibles sont localisés à la surface cellulaire (Yamaguchi et *al.*, 2009). Ce problème a été contourné par ajout d'une molécule d'ADNc. Les sélections sont faites avec des complexes molécules ARNm/ADNc-protéine. Avec ce système, les protéines contenant des ponts disulfures sont facilement repliées par changement de tampon. Cette approche a été validée par une sélection, contre le récepteur de l'interleukine 6 (Il-6R), d'une bibliothèque de peptides aléatoires (32 résidus). Cette approche a permis d'identifier des peptides ayant plusieurs ponts disulfures ce qui n'a pas été décrit auparavant avec d'autres techniques de sélection *in vitro*.

- *Compartimentation in vitro: Water in Oil Emulsions for Binding Selections (STABLE)*

Les méthodes de type *phage, ribosome ou ARN display* permettent de sélectionner efficacement des peptides ou des protéines ayant de nouvelles capacités d'interaction. Ces méthodes sont néanmoins plus difficiles à exploiter lorsque l'objectif est de faire émerger de nouveaux catalyseurs. Le problème fondamental est que le produit de la réaction enzymatique est généralement détaché de l'enzyme et libéré dans le milieu où il s'accumule d'autant plus rapidement que l'enzyme est performant. Dans la solution toutes les protéines, les plus comme les moins efficaces, sont mélangées. Ainsi, même s'il est possible d'associer chaque variant d'une bibliothèque d'enzymes à « son » gène, il n'est pas possible de savoir quel variant du catalyseur est le plus efficace si le lien avec les molécules qu'il a contribué à produire s'est évanoui par la diffusion du produit dans la solution. Ce problème peut être contourné de deux manières différentes : en liant l'enzyme à son substrat ou en créant de la

compartimentation, ce qui a été réalisé dans la nature avec l'apparition de la compartimentation cellulaire. Le concept central des approches de microcompartimentation est qu'une population d'enzyme pourra faire l'objet d'un processus de sélection à la condition que le produit accumulé par l'activité de chaque variant du catalyseur soit séquestré dans le compartiment qui contient ce variant ainsi que le gène qui le code.

En 1998, ce concept a été mis en pratique par Griffiths grâce à ses travaux visant la sélection du gène codant la HeaIII méthyltransférase à partir d'un mélange contenant 10^7 gènes différents codants pour d'autres enzymes (Tawfik and Griffiths, 1998). Les sélections ont été réalisées par un système de traduction-transcription dans des émulsions d'eau dans l'huile (Fig. 11).

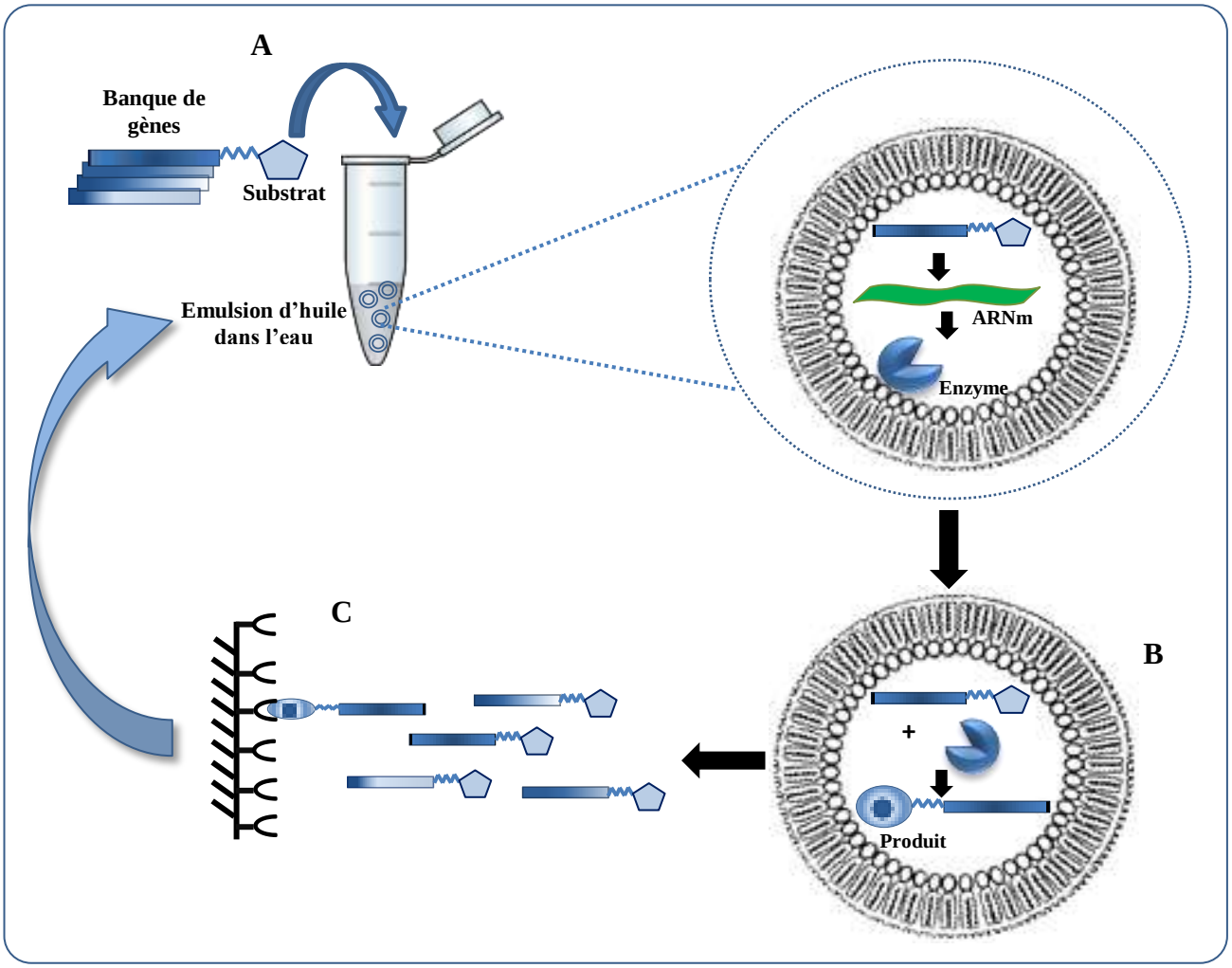


Fig. 11 : Schéma général d'un tour de sélection par compartimentation. A : Réaction de transcription/traduction couplées *in vitro* ayant lieu dans un milieu réactionnel localisé dans des émultions d'eau dans l'huile (1 émulsion = 1 gène = 1 enzyme). B : Les enzymes synthétisées transforment les substrats en produit qui restent liés aux gènes. C : Les produits obtenus par la réaction enzymatique sont récupérés par affinité permettant ainsi de collecter des gènes correspondant aux enzymes actives. Les gènes récupérés sont utilisés pour une analyse de séquences ou encore amplifiés pour être réinjectés dans un deuxième tour de sélection.

Des travaux intéressants concernant la sélection de protéines par cette technique ont été réalisés en 1999 par Yanagawa et *al.*. Cette équipe a décrit une méthode qui s'appuie sur une réaction de transcription/traduction au sein d'émulsions d'eau dans l'huile : en effet le polypeptide fusionné à la streptavidine est synthétisé puis lié à son ADN biotinilée. D'où la désignation « *STABLE : STA-biotin linkage in emulsions* ». Les molécules ADN-protéines récupérées des émulsions sont utilisées pour des sélections. La technique a été utilisée pour réaliser des sélections à partir d'une banque de Fabs randomisés en présence de billes magnétiques revêtues de fluorescéine avec deux températures de réaction différentes (température ambiante et 68°C) (Sumida, Doi, and Yanagawa, 2009). Les sélections à haute température ont permis d'isoler des Fabs plus stables que ceux identifiés à température ambiante mais moins stables que le type sauvage. Les affinités ont été mesurées et comparées au type sauvage : elles sont toutes du même ordre de grandeurs aux alentours de 2 nM.

Bien que les techniques d'évolution des protéines *in vitro* ont connu un grand succès ses dernières années, elles ont permis de réaliser des avancées considérables dans différents domaines thérapeutiques, biotechnologiques… Mais d'autres techniques ont aussi été développées : les techniques d'évolution *in vivo*. Elles ont également permis des avancées intéressantes et doivent donc prendre part à l'éventail des stratégies d'évolution possibles.

4.3. Les techniques de création de la diversité *in vivo*

Les méthodes d'évolution de protéines *in vivo* permettent la simplification du processus de mutagénèse et la procuration d'un environnement naturel et physiologique lors de la sélection du phénotype recherché. En effet, les paramètres intracellulaires, tels que la présence de modifications post-traductionnelles, le pH, la concentration en ions, peuvent influencer l'état fonctionnel des protéines. De ce fait, il est intéressant de générer de la diversité pour une protéine d'intérêt puis de sélectionner directement dans des cellules vivantes. Ceci peut être réalisé par deux approches différentes en fonction du type des cellules hôtes : les techniques d'évolution dans les cellules procaryotes ou dans les cellules eucaryotes (Blagodatski and Katanaev, 2011).

a) Les systèmes procaryotes d'évolution dirigée des protéines

Le système le plus communément utilisé par les biologistes est la bactérie *E.coli* du fait des connaissances disponibles, de la simplicité de sa manipulation et de la rapidité de sa

croissance. Cependant le taux de mutations naturel reste très insuffisant pour le but à atteindre. Ainsi des souches mutagènes ont été créées afin d'augmenter la fréquence spontanée de mutagenèse. Les souches les plus utilisées présentent une déficience dans la réparation de l'ADN telle que la souche commerciale XL1-Red d'*E.coli*. Cette souche est déficiente au niveau des gènes *mutD, mutS et mutT*. Suite à la transformation des bactéries par un plasmide contenant le gène d'intérêt, les mutations sont accumulées au cours de chaque cycle de réplication d'ADN avec une fréquence pouvant s'élever à 1 base pour 2000 nucléotides. Ceci est suffisant pour apporter un faible taux de mutations aléatoirement distribuées dans le gène cible et obtenir les changements voulus pour la protéine. Cette souche a été utilisée pour changer la spécificité de l'estérase de *Pseudomonas fluorescens* pour son substrat. En effet, on a pu obtenir des variants capables d'hydrolyser un 3-hydroxy ester stériquement encombré. Le criblage a été réalisé en faisant pousser les bactéries XL1-Red, exprimant l'estérase de *Pseudomonas fluorescens,* sur un milieu solide minimum contenant des indicateurs colorés (Bornscheuer, Altenbuchner, and Meyer, 1999). Les colonies ayant l'estérase mutée sont détectées par leur coloration rouge causée par l'augmentation du pH suite à l'hydrolyse du 3-hydroxy ester et la libération d'acide.

D'autre souches plus mutagènes ont été développées telles que la souche d'*E.coli* exprimant des ARN antisens, après induction à l'IPTG, provoquant ainsi un « *silencing* » simultané des gènes impliqués dans la réparation et la synthèse de l'ADN : mutS, mutD et ndk. Ceci conduit à une augmentation de 2000 fois de la fréquence des mutations spontanées (Nakashima and Tamura, 2009).

Ces souches ne permettent pas de cibler les gènes d'intérêt autrement qu'en propageant un plasmide puis en extrayant celui-ci, mais le contexte génétique est lui aussi susceptible d'être altéré. Une autre approche a alors été développée dans le but de muter préférentiellement l'ADN cible. Elle est basée sur l'utilisation d'une souche d'*E.coli* mutée sur la séquence du domaine structural de l'ADN polymérase I qui contribue à sa fidélité. Quand PolI est active lors de la réplication du plasmide ColE1, la polymérase, source d'erreurs, va affecter principalement la séquence du gène cible portée par le plasmide. Le taux de mutations peut atteindre $8.1*10^{-4}$ mutations par paire de base, 80000 fois plus que la fréquence naturelle. Les mutations peuvent s'étendre sur 3kb, avec une fréquence maximale au niveau des 700 premières paires de bases. Ce profil mutagène provient du rôle de Pol1 dans l'initiation de la réplication du plasmide, et les étapes suivantes de la réplication sont

dépendantes de PolIII qui est plus fidèle. Les mutations sont distribuées uniformément et sont plus fréquentes dans les cultures bactériennes maintenues en phase stationnaire. Ce système a permis d'augmenter de 150 fois la résistance bactérienne à l'aztreonam, chez les souches où le gène de la TEM-1 β-lactamase est ciblé (Camps et *al.*, 2003).

Toutes les techniques décrites ci-dessus ne permettent que l'évolution d'un gène à la fois. Mais récemment une approche intitulée « *Multiplex Automated Genome Engineering* » (MAGE), a été mise au point pour permettre de cibler et d'optimiser simultanément une série de gènes impliqués dans une voie de biosynthèse complexe (Wang et *al.*, 2009). La souche utilisée est la souche modifiée d'*E.coli* : EcNR2. Cette souche permet d'incorporer des oligonucléotides homologues au niveau des brins d'ADN tardifs lors de la réplication. Il est donc nécessaire d'introduire des oligonucléotides homologues au gène cible mais comportant des nucléotides substitués, au cours de la culture. Globalement, cette technique permet d'effectuer $4.9*10^8$ mutations par cycle et elle a été appliquée sur la voie de biosynthèse du 1desoxy-D-xylulose-5-phosphate (DXP), chez *E.coli*, dans le but d'augmenter la production industrielle de l'isoprenoide lycopéne. En effet au bout de 3 jours de culture bactérienne, des souches hyperproductrices ont été isolées avec 5 fois plus de production. L'analyse des séquences a permis d'observer que 24 gènes impliqués ont été mutés.

Pour conclure, ces techniques, impliquant des microorganismes procaryotes, sont intéressantes grâce à la simplicité des conditions de culture et le temps de génération réduit. De plus ces méthodes présentent un mode de sélection ou de criblage simple. Il est en effet basé soit sur la capacité des clones d'intérêt à résister à un antibiotique soit sur leur aptitude à produire une coloration. Cependant, elles sont limitées au niveau du type de gènes ciblés en particulier quand il s'agit de gènes de protéines eucaryotes exigeantes pour les conditions de repliement et les modifications post-traductionnelles d'où le recours à des cellules hôtes eucaryotes.

b) Les systèmes eucaryotes d'évolution dirigée des protéines

Le système eucaryote le plus simple et le plus connu est la levure. Une des avancées récentes utilisant la levure est la modification simultanée d'un gène cible en plusieurs sites par recombinaison multiples avec une bibliothèque d'oligonucléotides synthétiques diversifiés (Pirakitikulr et *al.*, 2010). Les cellules sont co-transformées avec d'une part un mélange d'oligonucléotides comprenant des régions complémentaires à la séquence cible et d'autre

part un plasmide linéaire contenant le gène cible. De manière générale, cette technique ressemble à la technique MAGE chez les bactéries. Elle a été testée pour la réversion d'un ou plusieurs codons non-sens dans les sélections classiques des marqueurs TRP1, ce qui a permis d'obtenir des clones positifs basés sur leur capacité de croître sur des milieux dépourvus de Tryptophane. La technique peut être avantageusement utilisée quand la modification combinatoire de plusieurs sites d'un gène cible est recherchée.

Une approche simple pour l'évolution dirigée des protéines *in vivo* est celle qui utilise des cellules incapables de réparer les mésappariements au niveau de l'ADN. Mais ce système reste non spécifique : il permet de « randomiser » le génome entier sans cibler le gène d'intérêt. Afin de remédier à ce problème, des stratégies inspirées de mécanismes naturels de l'amélioration de l'affinité des anticorps par hypermutation somatique et conversion génique ont été développées.

L'hypermutation somatique est un mécanisme qui a lieu lors de la mise en place de la réponse immune, suite à l'activation des cellules B et leur rencontre avec l'antigène. Ce processus est hautement spécifique. Il résulte d'une augmentation de la fréquence des mutations ponctuelles au niveau des séquences qui correspondent aux régions hypervariables des chaînes lourdes et légères des immunoglobulines. Dans ce cas, le taux de mutations s'élève à 10^{-3} par paire de base par génération : 10^{6} fois plus élevé que le taux de mutation dans le reste du génome. Ce processus nécessite l'activation de la cytidine désaminase inductible (AID) dont l'activité consiste en la désamination des résidus deoxycytidine des gènes en transcription. Les mutations sont introduites lors du mécanisme de réparation des dommages engendrés par l'AID. Le premier essai d'application de l'hypermutation somatique pour l'évolution dirigée de protéines non-immunoglobulines, a été réalisé sur la protéine fluorescente rouge monomérique (monomeric red fluorescent protein mRFP) qui a été exprimée dans une lignée de cellules B humaine, Burkitt lymphoma Ramos cells. Ces cellules en culture mutent continuellement le gène IgV (Sale and Neuberger, 1998). Le taux des mutations est dépendant de la région où le transgène est intégré : le taux le plus élevé est détecté quand le transgène d'intérêt est intégré à proximité du locus hypermutable de la chaîne lourde des Ig.

Il existe un autre mécanisme observé dans les cellules B pour générer de la diversité dans les gènes d'immunoglobulines. Ce mécanisme correspond à la transformation de pseudogènes existant naturellement chez les oiseaux et certains mammifères. Il prend place en

même temps que l'hypermutation somatique sous le contrôle de la même enzyme clé : AID. La lignée de cellules B de poulet DT40 est une excellente plateforme pour l'évolution dirigée des protéines *in vivo*, par comparaison à la même lignée humaine ou de souris. En effet ces cellules sont dotées des deux mécanismes mutagènes (hypermutation somatique et conversion génique) et d'un taux élevé d'intégration ciblée des transgènes. La sélection à partir d'une librairie diversifiée autonome (ADLib : autonomously diversifying library) a conduit très rapidement à l'obtention de différents clones qui produisent des IgM monoclonaux avec une efficacité et une affinité suffisamment importante pour réaliser des essais immunologiques (Seo et *al.*, 2006). Cette approche conduit à la production *ex vivo* de mAbs reconnaissant des protéines conservées au cours de l'évolution en contournant les problèmes d'immunogénicité rencontrés avec d'autres méthodes telles que les hybridomes. Les anticorps, ainsi obtenus sont aussi exploités pour des étapes ultérieures de *design* rationnel. D'autres essais ont été réalisés sur des protéines autres que les anticorps [(Burritt et *al.*, 1995) (Kanayama et *al.*, 2006)]. A titre d'exemple, l'équipe d'Ohmori a réussi à convertir la fluorescence bleue de cellules DT40 exprimant la BFP (Blue Fluorescent Protein) en une fluorescence verte par des évènements de conversions géniques. Ces derniers ont eu lieu entre le gène de la GFP (gène donneur) et celui de la BFP (gène accepteur) qui sont insérés au niveau du locus des IgL. Cette conversion génique est catalysée par l'AID, qui est exprimée dans les cellules DT40 et elle a conduit à l'insertion de fragments du gène de la GFP dans les zones homologues du gène de la BFP comme une sorte de *DNA shuffling in vivo.* Des travaux ultérieurs impliquant aussi le *design* rationnel ont permis d'obtenir des variants de GFP qui sont 3 fois plus fluorescents. La lignée DT40 présente l'avantage, par rapport aux autres lignées des cellules B (Rmos ou 18-81), d'insérer le gène d'intérêt de manière ciblée au niveau du locus de la chaîne légère des Ig, mais aussi de contrôler le processus de mutation en modulant l'expression de l'AID.

L'avantage des techniques citées ci-dessus est la simplicité de leur mise en œuvre. En effet, les cellules sont tout simplement mises en cultures et les pools d'ADN représentant des variants mutés sont par la suite récupérés puis analysés pour déterminer les mutations et les évènements géniques survenus.

Pour nos travaux nous nous sommes intéressés à une technique de sélection et de criblage particulière : *phage display* qui est détaillée dans la partie suivante de l'introduction.

C. Exposition sur phages M13

Dans notre travail, nous avons exploité uniquement l'exposition de nos protéines sur phages M13. D'une part parce que c'est la méthode de sélection qui est la mieux maitrisée au sein de notre équipe et d'autre part parce qu'il s'agit d'une technique relativement bien adaptée à la sélection rapide pour différentes cibles, dès lors qu'une bibliothèque de bonne qualité est disponible. Nous développerons, dans la section suivante, différents aspects de cette technique.

Quels types de polypeptides exposés sur phage ?

Plusieurs types de polypeptides ont été exposés sur phages. Depuis les bibliothèques de peptides linéaires aléatoires de 6 à 43 résidus [(Burritt et *al.*, 1995), (Haaparanta and Huse, 1995)] jusqu'aux bibliothèques de peptides exposés sous forme de boucle obtenus par des pont S-S (McConnell et *al.*, 1996).

Pour les chaînes polypeptidiques exposées en boucle sur phages, grâce à l'insertion de deux paires de deux résidus cystéines autour du gène d'intérêt, elles comportent des ponts disulfures croisés. Ces structures assez contraignantes sont efficacement exposées sur les phages et sont même utilisées pour la sélection contre des cibles ne tolérant pas les peptides linéaires. De plus, il a été décrit que la formation des ponts va améliorer l'affinité des ligands vis-à-vis des cibles (McLafferty et *al.*, 1993). Depuis ces travaux, les chercheurs ont concentré leurs efforts sur le moyen le plus efficace pour l'exposition de structures plus complexes comme les immunoglobulines et d'autres ossatures protéiques, dont le potentiel de reconnaissance est nettement plus riche que celui des peptides.

Un anticorps exposé sur phage : comment faire?

D'un point de vue structural, les anticorps sous la forme d'immunoglobulines G (IgG) sont formés de deux chaînes légères à deux domaines chacune : un domaine constant (CL) et un autre variable (VL). Ils comprennent aussi deux chaînes lourdes composées d'un domaine variable (VH) et de trois domaines constants chacune (CH). Les domaines VH et VL sont responsables de l'interaction anticorps-antigène alors que les domaines constants sont médiateurs de l'élimination des antigènes. Les anticorps entiers ne sont pas exprimés efficacement dans les bactéries, et par conséquent ne sont pas exposés efficacement sur phages. L'organisation en domaines de la structure des anticorps a toutefois permis

l'expression de différents fragments (Fig. 12) capable de se replier plus efficacement que la molécule entière en système procaryote. L'expression périplasmique utilisée pour l'exposition sur phage permet la formation des ponts disulfures et l'expression d'une partie significative de la population de fragments sous forme fonctionnelle à la surface de bactériophages. Le clonage des gènes (VL+CL+VH+CH1) a produit les fragments Fab (*Fragment antigen Binding*).

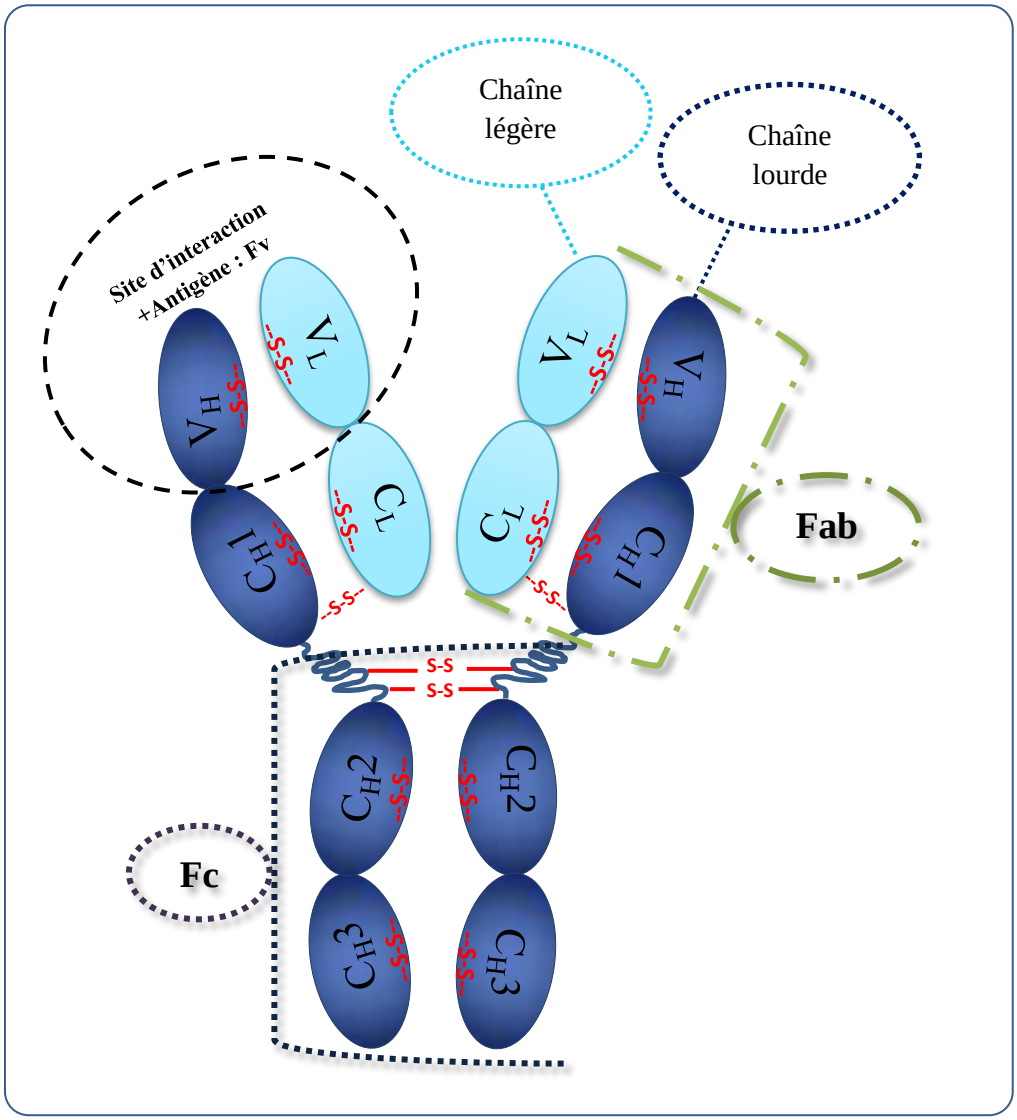


Fig. 12 : Schéma général d'un anticorps et les différents domaines qui le constituent.
-S-S- : ponts disulfures

Les deux fragments d'anticorps, Fv et Fab, peuvent être exprimés séparément : fusionné à la protéine de l'enveloppe du phage ou secrété dans le périplasme. L'environnement oxydant de ce dernier va permettre l'assemblage et le repliement des fragments Fab à la surface du phage. Les 2 domaines VL et VH ne s'associent pas de façon stable lorsqu'ils sont isolés mais ils peuvent être connectés par une séquence de liaison

(*linker)* flexible et forme alors de façon intramoléculaire des interactions entre les domaines V_H et V_L : ce type de construction, très utilisé en *phage display*, est appelé Fragment variable à simple chaîne ou scFv (pour *single chaine variable fragment*).

Comment recréer de la diversité dans les bibliothèques exprimées sur phages ?

L'obtention de ligands, avec de haute affinité, nécessite en général une bibliothèque très diverse dès le départ. Au niveau des acides nucléiques, la création de bibliothèques complexes et diverses, est réalisable grâce à toutes les techniques de mutagenèse et de recombinaison (*epPCR*…). Toutefois l'étape de transformation de l'information génétique dans les cellules hôtes reste une étape très contraignante : conserver la diversité lors de l'exposition sur phages reste une étape délicate.

Pour les anticorps !

Pour les anticorps, par exemple, la création de la diversité peut être effectuée par combinaison des chaînes hypervariables en particulier les VL et les VH. Waterhouse et ses collaborateurs ont décrit une approche permettant d'augmenter *in vivo* la diversité des banques (Waterhouse et *al.*, 1993). Ce système repose sur la recombinaison spécifique utilisant le site Cre-*lox* du phage P1. Dans ce système, les deux chaînes lourde et légère des fragments d'anticorps sont codées séparément : une sur le phagemide et l'autre sur le génome du phage. Ces gènes sont insérés entre 2 sites *loxP* hétérologues. Ainsi, dans une bactérie ayant le phagemide et infectée par le phage P1 recombinant, l'expression de la Cre-recombinase va déclencher une recombinaison des séquences VH et VL, pour produire des gènes réarrangés dont un codera pour un fragment Fab, potentiellement fonctionnel.

Des améliorations ont été introduites à ce système, la plus intéressante est l'introduction d'une pression de sélection tel que les gènes de résistance à un antibiotique permettant de sélectionner les phages ayant la bonne recombinaison (Geoffroy, Sodoyer, and Aujame, 1994).

Autres que les anticorps !

Une autre approche a été décrite par Collins intitulée *cosmix-plexing* (Collins et *al.*, 2001) permettant de créer de la diversité dans des peptides. Cette technique nécessite la construction d'une première bibliothèque de phagemides codant pour des polypeptides

dégénérés. Ces polypeptides sont subdivisés en plusieurs fragments séparés par des sites de restriction non-palindromiques. Le phagemide contient également un autre site de restriction unique appelé site de résolution. La bibliothèque ainsi construite, sera utilisée pour des sélections contre des cibles choisies dans le but d'enrichir la population en variants de faible affinité. Les gènes de ces variants ont été collectés puis recombinés *in vitro* pour recréer de la diversité. Il en résultera une nouvelle bibliothèque plus diverse et recelant potentiellement des molécules plus affines. Celle-ci a été soumise à d'autres tours de sélections. En général, les ligands obtenus par cette méthode ont des affinités comparables à celles des anticorps.

L'amélioration des affinités suite à la sélection des phages peut être aussi effectuée en mutant spécifiquement certains résidus des polypeptides obtenus. Pour cela, les séquences collectées lors du 1er tour de sélection sont analysées puis comparées par alignement de séquences. L'analyse des alignements de séquences va permettre d'identifier des résidus conservés et d'autres variables. Les résidus conservés pourraient être plutôt impliqués dans le repliement alors que ceux qui sont variables d'une séquence à une autre pourraient être impliqués dans les interactions. A partir de cette hypothèse, la mutation des résidus variables, potentiellement impliqués dans les interactions, suivie d'un second tour de sélection peut être une stratégie de maturation d'affinité (von Schantz et *al.*, 2009). D'une manière plus aléatoire, Thie et *al.* ont introduit par *error-prone PCR* des mutations aléatoires aux variants issus du 1er tour de sélection. Des mutants à haute affinité ont été alors isolés à l'issue d'un second tour de sélection réalisé dans des conditions très sélectives (Thie et *al.*, 2009).

Grover et son équipe ont procédé d'une autre façon pour des sélections, à partir d'une première banque de peptides contre des isoformes de canaux calciques de la membrane plasmique (plasma membrane Ca^{2+} pump : PMCA4) qui sont au nombre de 4. Elles ont abouties à l'identification de plusieurs interacteurs dont un bien particulier. Ce dernier inhibe, à la fois, 2 isoformes de PMCA et les constantes d'inhibition sont de 46 µM et de 105 µM. L'idée est alors de construire une banque basée sur ce peptide, une autre bibliothèque, à mutations limitées, dont les peptides différent du peptide original d'1, 2 ou 3 résidus. Elle a permis l'obtention d'un peptide « deuxième génération » qui inhibe les 4 isoformes de PMCA avec des K_i plus importantes qui varient entre 2.3 µM et 67 µM (Pande et *al.*, 2008).

Comment procéder à la sélection par phage display ?

Généralement, les méthodes de sélections sont basées sur l'affinité des peptides ou des protéines exposées sur phages pour une cible immobilisée sur un support (*biopanning*) (Pande, Szewczyk and Grover, 2010). Ceci implique la réalisation de certaines étapes :

- ***L'immobilisation de la cible*** : Si la cible est une protéine, cette étape se fait sur support en polystyrène, par simple adsorption. Les molécules non retenues à la surface sont éliminées par lavage et les sites non couverts par la cible sont bloqués par des protéines neutres (BSA) ou des détergents nonioniques (Tween). L'immobilisation de la cible peut également être obtenue par greffage covalent des protéines cibles sur un support solide comportant des fonctions réactives (époxydes, ou NHS esters par exemple) ou encore par immobilisation grâce à une étiquette « *tag* » ou une biotine sur un support préalablement greffé à la streptavidine ou avec un réactif spécifique du « *tag* ».
- ***Fixation des phages :*** les phages exposant les peptides de la banque dégénérée sont ajoutés à la cible immobilisée.
- ***Elimination des phages non fixés :*** le premier tour de sélection est, en général, réalisé dans des conditions modérément sélectives (par exemple sans lavages trop poussés) de façon à ne pas risquer de perdre des variants qui ne sont encore que peu représentés dans la population. Mais, plus on avance dans les tours de sélection plus cette étape peut être prolongée pour sélectionner préférentiellement les phages ayant une très bonne affinité.
- ***Elution des phages :*** cette étape peut être réalisée de deux manières différentes :

  ✓ Elution dans des conditions déstabilisantes (pH dénaturant acide ou basique): les liaisons entre l'interacteur et la cible peuvent souvent être rompues dans des conditions acides (pH = 2 à 2.5). Les phages étant résistants aux pH acides ne sont pas inactivés si l'incubation n'est pas prolongée.

  ✓ Compétition avec des molécules ayant de l'affinité pour la cible.

Par la suite les phages élués sont amplifiés par infection de cellules hôtes.

Ces étapes sont nécessaires pour réaliser un tour de sélection en *phage display*, mais plusieurs modifications peuvent être introduites, en particulier, au niveau de l'immobilisation

de la cible et de la procédure d'élution. L'immobilisation des cibles peut se faire au moyen des étiquettes (Koide et *al.*, 2009) ou des anticorps permettant, ainsi, à la cible de maintenir sa conformation naturelle et lui assurer une surface accessible maximale. Le système le plus communément utilisé est l'interaction biotine/streptavidine ou avidine. En effet, la cible biotinylée est capturée par la streptavidine qui couvre la surface, soit de la plaque, soit de billes magnétiques (Park, Cropek, and Banta, 2010). D'autres cibles non protéiques nécessitent la mise au point de méthodes particulières. Les haptènes, par exemple, sont immobilisés par le biais de leur conjugués (Sheedy, MacKenzie and Hall, 2007). De plus, une étape de sélection négative, sur plaque ou billes non couvertes par la cible, peut être envisageable pour minimiser la fixation non-spécifique de phages.

D. Les ossatures protéiques alternatives aux anticorps

1. Pour quoi une alternative aux anticorps ?

Les anticorps ont, longtemps, été la seule option envisageable lorsqu'il est nécessaire de disposer d'une protéine capable d'interagir avec une molécule cible choisie *a priori*. Même si depuis plus de dix ans, des protéines alternatives aux anticorps ont été proposées, ces derniers sont encore, et de très loin, les interacteurs spécifiques les plus utilisés.

Des anticorps hautement spécifiques et affins peuvent être obtenus par diverses méthodes : ils peuvent être directement produits par le système immunitaire d'un organisme immunisé, par des cellules du système immunitaire immortalisées, ou identifiés au sein de bibliothèques d'anticorps artificiellement reconstituées. Les anticorps présentent l'avantage de reposer sur une architecture exceptionnellement versatile : ils sont capables de se lier spécifiquement et avec une forte affinité à des macromolécules comme des protéines mais aussi avec des sucres, des peptides et diverses autres molécules… Du point de vue des applications pharmaceutiques, ce sont des molécules qui peuvent être générées à partir d'une ossature presque entièrement humaine, et qui sont donc potentiellement moins immunogènes qu'une protéine non humaine. Enfin, sous forme IgG, leur taille leur confère une longue période de demi-vie dans le sérum, ce qui est une propriété importante pour certaines applications thérapeutiques. En revanche, ces protéines sont formées par un assemblage complexe de chaines polypeptidiques multidomaines associées entre elles par des ponts disulfures. De plus, ces molécules sont naturellement glycosilées dans leur forme fonctionnelle. On conçoit donc que leur expression dans des systèmes d'expression

procaryotes ne soit que très peu efficace. Leurs propriétés biophysiques et leur organisation moléculaire complexe rendent leur production souvent difficile et coûteuse (de Marco, 2009). Si les applications thérapeutiques des anticorps ont connu des développements très importants, leur production repose toujours sur des systèmes d'expression en cellules animales en culture. Si de tels systèmes d'expression sont possibles pour la production de molécules à très haute valeur ajoutée, cette contrainte rend plus difficilement envisageable d'autres applications pour lesquelles le coût et l'efficacité des systèmes d'expression procaryotes seraient nettement plus appropriés.

L'usage d'un anticorps entier pour des applications thérapeutiques peut également poser des problèmes reliés à sa taille lorsqu'il est question de pénétration tissulaire. Des molécules dérivées d'anticorps, plus petites, sont alors utilisées : les fragments Fab, les scFv, les mini-anticorps multivalents. Qu'il s'agisse d'anticorps entiers ou de fragments, leur repliement suppose l'établissement de pont S-S dont la formation est défavorisée dans les milieux réducteurs tel que le cytoplasme des cellules hôtes. De plus, les fragments d'anticorps ont souvent une tendance prononcée à l'agrégation, cet effet étant encore aggravé s'ils sont fusionnés à d'autres domaines ou fragments effecteurs eux-mêmes sensibles à l'agrégation.

Les anticorps sont donc des molécules extrêmement utilisées, mais leurs utilisations pourraient être plus larges si elles n'étaient sérieusement contraintes par des propriétés de repliement inadaptées à certains environnements ou à certaines conditions. C'est en particulier le cas lorsqu'on cherche à exprimer un anticorps dans une bactérie sans emprunter les voies de sécrétion ou dans le compartiment cytosolique d'une cellule eucaryote. Développer les moyens de créer et de mettre en œuvre des anticorps actifs dans l'environnement intracellulaire, souvent appelés *intrabodies* ou intracorps est devenu un axe de recherche en soi (Moutel and Perez, 2009). L'objectif général serait de pouvoir disposer d'anticorps ayant la capacité de reconnaître une protéine intracellulaire cible dans son environnement au sein d'une cellule vivante permettant ainsi de suivre sa dynamique. Il est également possible d'utiliser l'intracorps pour interférer avec la fonction de la cible en empêchant son interaction avec ses partenaires dans le but par exemple d'élucider un processus hypothétique ou encore pour valider une cible pharmacologique. Par exemple, différents « *intrabodies* non-perturbateurs » fluorescents ont été développés par Perez et ses collaborateurs. Ceux qui ont été dirigés contre les formes liées au GTP de la tubuline, par exemple, ont permis à cette équipe de proposer un modèle expliquant la dynamique instable des microtubules (Dimitrov

et *al.*, 2008). L'identification d'anticorps actifs, au niveau intracellulaire, a nécessité une double sélection : une sélection pour des scFv spécifiques d'une cible donnée, puis parmi ceux-ci, l'identification de scFv opérationnels dans des conditions d'utilisation intracellulaires. Une autre approche possible est de concevoir puis de construire une bibliothèque d'anticorps dont l'ossature est particulièrement robuste. Une telle ossature a par exemple pu être élaborée par évolution dirigée d'anticorps actifs par expression cytoplasmique chez *E. coli*. Il apparaît donc que si une sélection est spécifiquement orientée vers l'amélioration du repliement et de la solubilité, des ossatures immunoglobulines aux propriétés améliorées peuvent être identifiées. Ce type d'ossature peut alors être exploité pour construire des bibliothèques par substitution aléatoires des CDR. L'expression de ce type d'anticorps a fait aussi l'objet de travaux de l'équipe de Martineau (Guglielmi et *al.*, 2011) montrant la nécessité de l'optimisation des conditions d'expression des *intrabodies* aussi bien dans *E.coli* que dans les cellules mammifères. Cette optimisation est possible directement dans les cellules humaines par le biais de la fusion à la GFP. Ce travail suggère également que, même si ce problème a rarement été considéré dans ce sens, une protéine évoluée pour s'exprimer favorablement chez une bactérie n'est pas nécessairement exprimée de façon optimale chez un hôte eucaryote.

Les *intrabodies* seront nécessaires pour développer certaines approches thérapeutiques par exemple pour former des immunotoxines utilisables comme anticancéreux ou encore pour acheminer des gènes thérapeutiques (Ahmad et *al.*, 2012). En effet fusionnés à des toxines, des nucléotides radioactifs ou certaines drogues, les *intrabodies* peuvent être considérés comme des immunotoxines qui sont délivrés spécifiquement à des cellules présentant des antigènes marqueurs du cancer. Plusieurs anticorps ont été développés pour ce type d'applications dont on peut citer l'immunotoxine formée par la fusion du fragment Fab humain anti-AChR à l'exotoxine A de *Pseudomonas*. Cet immunotoxine, une fois internalisée, va provoquer la mort des cellules cancéreuses. Nous pouvons citer aussi le fameux Certolizumab Pegol. Ce dernier est un immunoinhibiteur bloquant la TNF-α qui a été récemment validé, par la FDA, comme traitement pour la maladie de Crohn et pour l'arthrite rhumatoïde.

2. Des anticorps optimisés aux ossatures alternatives

Les protéines alternatives aux anticorps peuvent être aussi une stratégie pour contourner les limitations physicochimiques des domaines Ig tout en gardant les propriétés de

spécificité et d'affinité. Si il a, un temps, été envisagé que seule l'architecture de type anticorps était apte à être fonctionnalisée de façon à interagir avec n'importe quel antigène, il est désormais clair que de nombreuses autres architectures protéiques peuvent également être support d'interaction multiple, à la condition de recréer artificiellement un processus de diversification/sélection/amplification similaire à celui observé dans la réponse immunitaire.

L'objectif de ces travaux est donc de parvenir à des protéines qui comme les anticorps soient capables d'interagir spécifiquement avec n'importe quelle cible choisie *a priori*, sans pour autant présenter les inconvénients de ceux-ci : sensibilité aux environnements réducteur et tendance à former des corps d'inclusion. Une telle ossature alternative doit donc avoir des propriétés physico-chimiques améliorées. Plus précisément, les ossatures protéiques alternatives doivent remplir les attentes de stabilité thermique et chimique d'une part et de solubilité et résistances aux protéases d'autre part. Elles doivent se replier en une chaîne polypeptidique unique sans ponts S-S. Leurs séquences doivent être aussi tolérantes à la mutagénèse permettant ainsi la création de plus de la diversité. Les critères cités ci-dessus sont essentiels pour garantir une expression efficace de protéines fonctionnelles dans le cytoplasme cellulaire.

3. Quelles ossatures protéiques pour quelles applications ?

Plusieurs approches ont été proposées pour créer des protéines utilisables en tant qu'alternatives aux anticorps. Différentes classifications ont été proposées dans la littérature pour la cinquantaine d'ossatures décrites. Dans cette section, on passera en revue une partie de ces ossatures en citant les exemples les plus pertinents. Toutes les ossatures n'ont pas été poussées jusqu'au même stade de développement. Les moins avancées demeurent au stade de la proposition, tandis que les ossatures les plus avancées ont démontré leur potentiel, par exemple, pour les applications thérapeutiques et certaines sont même en essais précliniques et cliniques (Löfblom, Frejd, et Ståhl, 2011).

3.1. Les adnectines, ^{10}Fn3 ou encore « *monobody* »

Les adnectines constituent une des ossatures les plus étudiées sans doute pour leur potentiel thérapeutique. Cette ossature doit son nom au 10^{éme} domaine de la fibronectine humaine ^{10}Fn3. La forme naturelle de ce domaine comporte 94 résidus**.** Elle montre une similitude structurale avec les Ig naturels ; elle se replie en sandwich de feuillets β reliés par 6 boucles dont trois sont la cible de la diversification (Fig. 13).

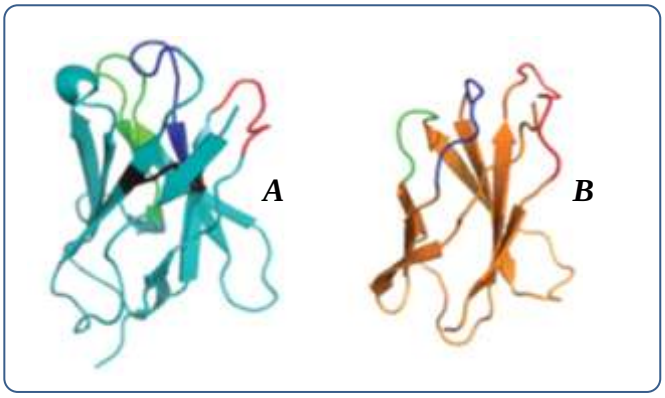


Fig. 13 : Schéma comparant la structure tridimensionnelle d'un domaine V_H d'un anticorps (A) (PDB ID: 1ITGY(Harris, Skaletsky, et McPherson, 1998) avec celle d'une ^{10}Fn3 (B): La similitude structurale réside dans l'existence chez l'adenectine de 3 boucles (en bleu, vert et rouge) qui sont similaire au 3 boucles CDR du V_H qui comporte un pont S-S (en noir).

Des bibliothèques synthétiques ont été construites, pour lesquelles les boucles équivalentes aux CDR des domaines Ig ont été rendues aléatoires. Des sélections, contre diverses cibles ont été réalisées par *phages*, *ribosomes* et même *yeast display* permettant ainsi d'obtenir des interacteurs dont les meilleurs montrent des affinités pour leur cible de l'ordre du nano et même du picomolaire (Lipovsek, 2011). L'une des molécules d'adnectine identifiée (CT-322) est spécifique au récepteur 2 du facteur de croissance vasculaire et endothéliale (vascular endothelial growth factor (VEGF) receptor 2 : VEGFR-2) (Getmanova et al., 2006). Ce variant a été obtenu par maturation d'affinité et de spécificité avec un interacteur original du VEGFR-2 murin.

In vivo, cette adnectine bloque l'activité du VEGFR-2 humain et murin à la fois et présente une action préventive de la croissance tumorale et de métastases dans deux modèles de tumeurs pancréatiques chez la souris. Plus récemment, l'adnectine a même montré une activité anti-angiogénique et antitumorale dans le modèle tumoral de la xénogreffe Colo-205. Les essais cliniques de la phase I témoignent d'une lente disparition dans le plasma avec une durée de demi-vie de 3 à 4 jours mais aussi l'absence de tout effet indésirable (un taux élevé de VEGF-A dans le plasma), une légère toxicité et une activité antitumorale prometteuse pour les doses maximales tolérées. La CT-322 a été conduite à des essais cliniques phase II comme unique agent dans un glioblastome multiforme ainsi qu'en complément à un traitement par chimiothérapie du carcinome des poumons et du colon (Tolcher et *al.*, 2011).

3.2. Les « *Affibodies* »

L'ossature «*affibody*» dérive du domaine B de la protéine A du staphylocoque. Cette protéine de 58 résidus se repliant en tonneau de 3 hélices α a été sujette à différentes méthodes d'ingénierie (J Löfblom et *al.*, 2010) (Fig. 14).

Fig. 14: Structure tridimensionnelle d'une molécule d'*affibody* (en rose) en complexe avec la région extracellulaire du récepteur 2 du facteur de croissance épidermique humain en bleu (Human Epidermal Growth Factor Receptor-2 : HER2) (PDB: 3MZW).

En effet, si cette ossature a été parmi les plus précocement décrites (Nord et *al.*, 1997), une ossature dite de 2^{éme} génération (Kronqvist et *al.*, 2011) a été obtenue en mutant 23 résidus des 58 acides aminés de la protéine sauvage dont 13, localisés sur la 1^{ére} et la 2^{éme} hélice, sont rendus aléatoirement variables. Des molécules d'*affibodies* ont été isolées suite à des sélections contre une large gamme de cibles montrant une grande stabilité thermique et structurale ainsi qu'une activité inhibitrice d'interaction protéine/protéine ou encore dans les systèmes d'acheminement des molécules actives. Très récemment, des travaux ont décrit la sélection par exposition sur phage et sur bactérie (sur staphylocoque) de nouvelles molécules d'*affibodies* contre le récepteur HER3 qui sont capables de bloquer la phosphorylation de ce récepteur.

Cette ossature a été aussi utilisée sous une forme modifiée avec un complexe contenant un isotope radioactif dans plusieurs essais d'imagerie de tumeurs. L'un des travaux les plus pertinents est le travail de Tolmachev qui porte sur la comparaison de deux traceurs radioactifs de l'expression du HER2 dans les cellules tumorales (Orlova et *al.*, 2009). En effet cette équipe a utilisé d'une part un anticorps monoclonal humanisé : le trastuzumab et d'autre part une petite molécule d'*affibody* (7 kDa) : $Z_{HER2:342}$ marquées à ^{124}I au niveau d'un linker *p*-iodobenzoate. Les 2 traceurs interagissent spécifiquement avec HER2 exprimé *in vitro* dans des cellules mais aussi *in vivo* dans des xénogreffes. Il est montré que le trastuzumab radioactif est internalisé et dégradé plus rapidement que la $Z_{HER2:342}$ ce qui permet une

meilleure rétention de la radioactivité délivrée par l'*affibody*. Ces résultats ont été confirmés par imagerie chez de petits animaux *ex vivo*. La molécule d'*affibody* a alors un meilleur contraste en imagerie de l'expression du HER2.

Une nouvelle génération d'*affibodies* (ABY-025) a été construite ; elle est saine et non-immunogène lors des tests précliniques et elle a atteint le stade des essais cliniques (Ahlgren et *al.*, 2010). Une autre équipe s'est intéressée à un autre type de marquage des *affibodies* par des sondes émettant dans l'infrarouge (Chernomordik et *al.*, 2010). Ces *affibodies* sont utilisés pour cibler des tumeurs et la quantification de l'expression du récepteur HER2. Toutes ses avancées montrent que ces molécules d'*affibodies* modulées sont des agents prometteurs pour la thérapie ainsi que pour l'imagerie *in vivo*.

3.3. Les anticalines

Les anticalines constituent une ossature dérivée des lipocalines naturelles structurées en tonneaux β contenant 4 boucles (Flower, 1996). Cette famille est observée chez différents organismes, depuis les bactéries jusqu'à l'homme. Elles sont impliquées dans le stockage et le transport de molécules hydrophobes, telles que le rétinol ou les stéroïdes, ou de molécules chimiquement instables, telles que les vitamines et les phérohormones. Les lipocalines présentent une partie structurée en tonneau β. Ce repliement rigide semble supporter des boucles de différentes longueurs, de séquences et conformations variables (Fig. 15) (Schlehuber et Skerra, 2005). Ces boucles sont exposées du côté de l'ouverture du tonneau rappelant la façon dont les anticorps exposent leurs boucles hypervariables. La *bilin-binding protein* (proteine de pigment bleu) de *Pieris brassicae* (papillon) est la première lipocaline qui a servi de modèle pour la création de nouveaux sites d'interaction à partir de cette ossature. En effet, 16 acides aminés, localisés sur les 4 boucles de l'ouverture du tonneau et des régions adjacentes aux feuillets β, ont été aléatoirement mutés. A partir de cette banque d' « anticalines » obtenue, des sélections ont été réalisées contre de petites molécules telles que la fluorescéine, la digoxigénine (stéroïde), des esters d'acides phtaliques et la doxorubicine (principe actif utilisé dans la chimiothérapie). Les affinités mesurées sont de l'ordre du nanomolaire ce qui semble être plus affins que les lipocalines pour leurs ligands naturels.

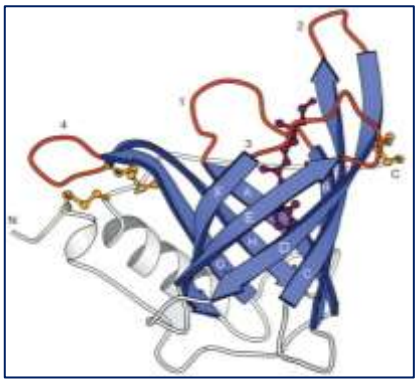


Fig. 15 : Structure tridimensionnelle de la «*retinol binding protein*» humaine en complexe avec le rétinol (PDB : 1RBP) (en magenta) : Cette structure typique d'une lipocalines montre une protéine constituée de 8 feuillets β très conservés dans cette famille (en bleu de A-H) avec 4 boucles variables (en rouge de 1-4), une hélice α, une boucle et 2 segments N et C-terminaux (en gris). Cette structure présente aussi 3 pont S-S (en jaune) (Schlehuber et Skerra, 2005).

Ces protéines ont été exposées sur phages et même sur *E.coli* (Binder et *al.*, 2010) grâce à un système d'auto-transport mettant en œuvre la serine protéase EspP de la membrane bactérienne. Ce système a été appliqué pour générer une anticaline immunostimulante qui interagit avec et bloque la région extracellulaire de la CTLA-4 humaine. Cette famille est exploitée par *Pieris protéolab*. Des résultats, exposés lors du colloque international des sociétés *BioIron* Mai 2011, portent sur les essais précliniques *in vitro* et *in vivo* d'une anticaline ciblant la hepcidine (the International BioIron Society Meeting) (http://www.pieris-ag.com). Cette dernière est une hormone peptidique sécrétée par le foie régulant l'homéostasie du fer dans l'organisme. L'effet antagoniste, avec haute spécificité et affinité de fixation, de l'anticaline à l'hepcidine pourrait traiter l'anémie.

3.4. Knottines ou les miniprotéines à nœud cystéine

Les knottines sont une famille de protéines de petite taille caractérisées par un repliement tertiaire très stable. En effet, ces protéines d'environs 30 acides aminés se replient à l'aide de 3 ponts S-S ce qui leur confère la forme d'un nœud cystéine (Sommerhoff et *al.*, 2010) (Fig. 16).

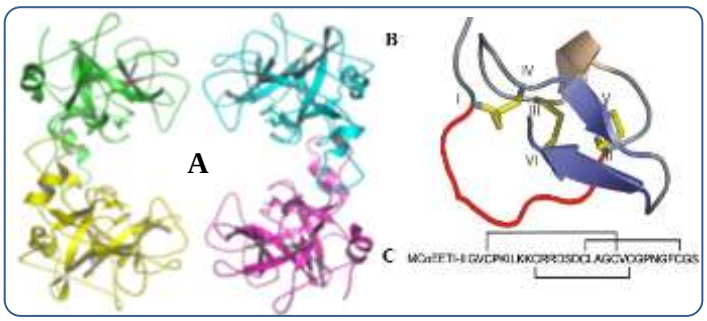


Fig. 16 : A : Structure tridimensionnelle de la Tryptase β des mastocytes humains : un tétramère de 4 miniprotéines à nœuds cystéines. B : Structure de la miniprotéine à nœuds cystéine MCoEETI-II : les chaînes latérales des cystéines (de I à VI) sont représentées en jaune avec la boucle inhibitrice en rouge formée entre Cys I et II. C : Séquence en AA de la MCoEETI-II avec les pont S-S qui se forment entre ses Cys.

Cette famille a été exploitée l'équipe de Cochran dans le but de développer des interacteurs knottines ayant des affinités de l'ordre du nanomolaire pour des intégrines par exposition sur levure (Kimura et *al.*, 2009). En effet, ces intégrines sont des récepteurs qui s'expriment dans les cellules tumorales ou dans les vaisseaux sanguins néoformés. Elles interagissent avec ses ligands par une séquence bien déterminée Arg-Gly-Asp (motif RDG). Des knottines ont été générées à partir d'un inhibiteur de trypsine d'*Ecballium elaterium* (*Ecballium elaterium tripsin inhibitor* : EETI-II). La boucle de cet inhibiteur de 6 acides aminés a été modifiée par une autre boucle contenant 11 acides aminés. La séquence de cette boucle contient le motif RDG à différentes positions et les autres résidus sont aléatoirement mutés. Les variants identifiés ont été par la suite marqués par un composé fluorescent (^{18}F) pour l'imagerie (Miao et *al.*, 2009) ou encore par un élément radioactif (^{177}Lu) (Jiang et *al.*, 2011) à leur extrémité N-terminale. Ces travaux ont permis un marquage efficace *in vivo*, ils ont aussi démontré une inhibition de l'agrégation des plaquettes.

Des travaux postérieurs ont aussi exploité des ossatures dont la structure dépend fortement de la formation des ponts S-S tels que les domaines humains *kringle*. En effet, Lee et ses collaborateurs ont construit un large répertoire de molécules dérivées des domaines humains *kringle* et ont isolé par la suite d'interacteurs spécifiques au TNF, DR5 et DR5 (Lee et *al.*, 2010).

3.5. Les affilines

Les affilines constituent une famille de protéines inspirée de l'architecture d'une protéine humaine : la γB cristalline. C'est une protéine de 176 acides aminés structurée en feuillets β (Ebersbach et al., 2007) dont l'homologue chez les bovins de 174 acides aminés, est présenté Fig. 17.

Fig. 17 : Représentation de la γ-B-cristalline bovine (PDB 1AMM) qui a été utilisée comme modèle pour l'architecture de la γ-B-cristalline humaine. Les positions 2, 4, 6, 15, 17, 19, 36 et 38 (en rouge) correspondent aux résidus utilisés pour la construction de la banque.

La γB cristalline est normalement exprimée dans le cristallin des vertébrés au cours de l'embryogénèse et elle est responsable des propriétés de réfraction du cristallin. La particularité de cette protéine réside dans le fait que, bien qu'elle n'ait aucun partenaire connu ni d'activité enzymatique bien déterminée, elle est dotée d'une stabilité exceptionnelle. Elle est par ailleurs d'une taille réduite et d'une solubilité faisant d'elle une ossature idéale (Ebersbach et *al.*, 2007) pour la création d'agents d'interaction intracellulaires (Mirecka et *al.*, 2009). En effet la création d'une bibliothèque de γB cristalline, puis des sélections par *phage display*, contre la protéine E7 du papillomavirus humain ont permis d'identifier des variants qui interagissent avec la cible à des affinités de l'ordre du nanomolaire. L'étape ultérieure d'expression intracellulaire, dans des cellules mammifères, a montré que l'affiline antiE7 inhibe la prolifération des cellules E7 positives par comparaison avec son expression dans les cellules E7 négatives ou encore celles exprimant la γ-B-cristalline sauvage. Ainsi les affilines semblent être des alternatives convenables pour des applications intracellulaires puisqu'elles sont fonctionnelles même dans le milieu réducteur des cellules mammifères.

3.6. Les protéines à motifs répétés

Les protéines à motifs répétés constituent une classe bien particulière et très intéressante parmi les ossatures protéiques jusqu'à lors décrites. Ces protéines à motifs structuraux répétés sont très répandues dans la nature et leur exploitation dans le domaine de création d'alternatives aux anticorps s'avère très pertinente en particulier depuis la description du système immunitaire de certains poissons sans mâchoires (lamproies et myxines). En effet, le système immunitaire de ces vertébrés primitifs semble avoir évolué différemment des vertébrés à mâchoire (Alder et *al.*, 2005). Les récepteurs lymphocytaires variables des lamproies, VLR (*Variable Lymphocyte Receptors*) sont constitués non pas de domaines immunoglobulines mais de modules répétés de type *Leucines Rich Repeat* (LRR) (Velikovsky et *al.*, 2009). Cette même ossature est aussi observée dans plusieurs récepteurs immuns innés tels que les *Toll-like receptor*, ou encore chez les protéines de résistance aux maladies chez les plantes. Sur le plan évolutif, les VLR sont les récepteurs de l'immunité adaptative les plus anciens. Ce sont les seuls récepteurs naturels d'antigènes à avoir une ossature différente de celle des Ig. Les VLR des lamproies, à titre d'exemple, sont produits dans les lymphocytes par des évènements de réarrangements génétiques somatiques des modules LRR diversifiés. Ils sont aussi constitués d'un motif N-terminal (LRRNT) suivi d'un premier motif LRR1 puis jusqu'à 9 motifs LRR répétés à 24 résidus aléatoires (LRRV) puis un motif LRR final

(LRRVe), un peptide de connexion (CP) et un motif C-terminal (LRRCT) (Fig. 18) (S.-C. Lee et *al.*, 2012). Ils présentent, donc, naturellement les propriétés de protéines alternatives naturelles aux anticorps et sont donc des candidats de choix pour des applications telles que les biocapteurs, la bio-imagerie et la purification.

Fig. 18 : Structure générale d'un VLR de lamproie montrant qu'il est constitué de : N-terminal (LRRNT) - LRR1- $(LRRV)_n$ - LRRVe – LRRCP - LRRCT. Où n= 1 à 9 (S.-C. Lee et al. 2012).

a) Les protéines à motifs répétés quelle origine génomique ?

Ces protéines à structure particulière, formées par l'assemblage de motifs répétés, semblent provenir, sur le plan génétique, d'évènements de recombinaison et de duplication intragéniques. Les séquences répétées provenant des protéines de ces familles s' « autoassemblent » pour générer des arrangements réguliers soit linéaires soit sous forme d'une superhélice (Fig. 19).

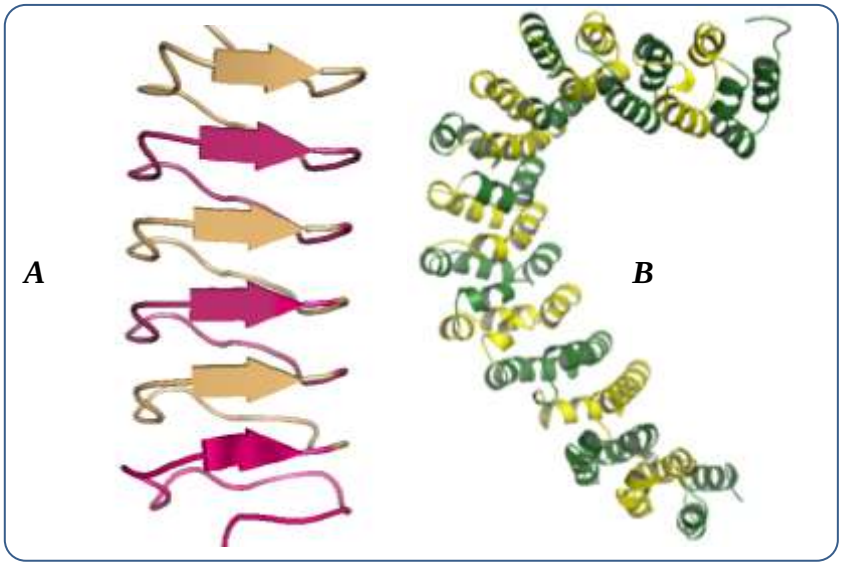


Fig. 19 : Structure tertiaire de 2 protéines à motifs structuraux répétés. A : La protéine antigel d'insecte (*Tenebrio molitor*, dans la PDB : 1ezg) a une structure linéaire exclusivement β. B : La phosphatase 2aPR65/A (*H. sapiens, dans PDB : 1b3u*) présente une structure en solénoïde formée par la répétition de motifs en hélices α.

Théoriquement, le nombre de modules qui peuvent être insérés dans une protéine est illimité tant que ça n'engendre pas d'encombrement stérique. L'arrangement des motifs répétés permet d'avoir une structure très stable avec une large surface exposée au solvant,

susceptible d'interagir avec des partenaires. Le développement de ce type de protéines semble être une manière simple avec laquelle les organismes élargissent leur répertoire de fonctions cellulaires. De telles protéines ont des capacités d'interactions exploitées dans la cellule pour des fonctions diverses par exemple de transport, d'assemblage des complexes, de régulation. Si la capacité de générer des protéines à motifs répétés est commune à tous les organismes, ce type de protéines est plus fréquent chez les eucaryotes que chez les procaryotes, et chez les métazoaires en particulier (Andrade, Perez-Iratxeta, et Ponting, 2001). Ceci peut être associé à l'augmentation de la complexité des fonctions cellulaires plus facilement accessible par assemblage de répétions à partir du génome préexistant.

b) Quelles sont les différentes familles de protéines à motifs répétés et où sont-elles observées dans la nature ?

Grâce au séquençage de génomes variés, la collecte des séquences et le développement des techniques d'analyse de séquences, différentes familles de protéines à motifs répétés ont été décrites. Les 6 familles, les plus importantes, sont ici classées selon la structure secondaire de leur motif ou « *repeat* » en spécifiant leurs propriétés et leur distribution dans les protéines naturelles.

i. Protéines à motif exclusivement β

- *Les β-propellers*

Cette structure particulière est observée chez différentes protéines couvrant une large gamme de fonctions : à titre d'exemple dans la chaîne lourde de la méthylamine déshydrogénase (PQQ repeats), le régulateur de la condensation chromosomique (RCC1 repeats) et la galactose oxydase (Kelch repeats) et dans la neuraminidase virale qui est la première structure résolue de cette famille (Varghese et *al.*, 1992). Au niveau fonctionnel, cette famille comporte des protéines à activité enzymatique est d'autres sans activité catalytique mais une capacité d'interaction avec des ligands de nature variée selon les protéines. Cette structure a été identifiée, pour la première fois, dans la sous unité β de la protéine G (Fong et *al.*, 1986) qui reste la famille de protéine à motif β-propeller la plus étudiée et dont la structure a été résolue en 1996 (Garcia-Higuera et *al.*, 1996). A partir de cette structure, ce motif a été décrit comme étant constitué de 40 acides aminés présentant des résidus particulièrement conservés : un Trp (W) et une Asp (D) d'où la désignation de motif

WD-40 (Fig. 20). Il s'organise en 4 brins β formant des courts feuillets antiparallèles. Ces feuillets sont disposés autour d'un axe central (Smith et *al.*, 1999). Le nombre de répétitions est variable de 4 à 8 mais cela forme toujours une structure circulaire, où le dernier motif de la séquence est au contact du premier. L'arrangement structural de la répétition de ce motif avec une pseudo-symétrie axiale lui confère une forme d'hélice de bateau (*propeller*).

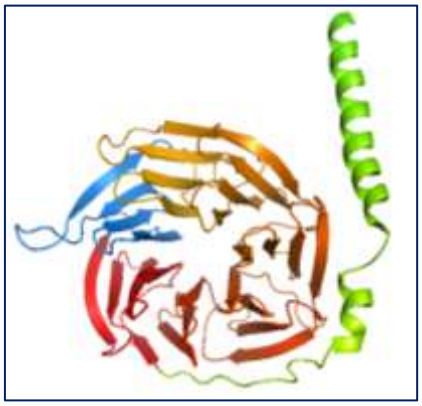


Fig. 20 : Structure d'une protéine à motif WD : la Gβ (PDB : 1gp2).

Au niveau du gène, la répétition structurale ne correspond pas exactement à une séquence d'ADN répétée. Mais dans ce cas, le motif répété correspond aux 3 premiers brins d'une pâle et au dernier brin de celle qui la suit. Cette interaction particulière contribue au verrouillage de la structure sur elle-même : par interaction entre un brin du dernier motif avec le feuillet de la première pâle. Cette structure, en forme de cône évasé, comporte 3 surfaces d'interactions : le sommet, le fond et la surface extérieure. Cependant, l'observation des structures résolues pour cette famille de repliement a permis de conclure que l'interaction se fait préférentiellement, tout au long de l'axe de la protéine, avec la surface formée par des N-terminaux des feuillets β intérieurs constituant ainsi une sorte de « *supersite* » d'interaction. L'intérieur de la protéine semble être très étroit pour former des interactions protéines-protéines, il est plutôt convenable pour de petites molécules et joue un rôle essentiel dans le maintien de la structure de la protéine.

- *β-Trefoils*

Cette famille de protéine est basée sur un motif répété en feuillet β. Les structures β-Trefoils ont été observées en premier dans le facteur de croissance des fibroblastes. Elle est constituée de 6 paires de feuillets β antiparallèles en épingle à cheveux : 3 paires vont former un tonneau et les 3 restantes s'organisent en une sorte de « *cap* » triangulaire. Cette ossature est constituée de 3 répétitions structurales qui ne sont pas détectables au niveau de la séquence (Ponting et Russell, 2000) (Fig. 21), probablement parce que la divergence des séquences

depuis la duplication a effacé les similitudes entre ces modules. Cependant, il a été, récemment, démontré qu'il est possible de refaire artificiellement le chemin inverse de ce processus évolutif probable et de récréer, par « déconstruction symétrique », une protéine *β-Trefoil* composée de motifs identiques (J. Lee et *al.*, 2011).

Fig. 21 : Structure typique d'une *β-Trefoil* : Facteur de croissance des fibroblastes.

Contrairement aux *β-propellers*, les *β-Trefoils* ne semblent pas posséder un « *supersite* » pour l'interaction avec les ligands. En effet, les membres de cette famille présentent différentes localisations pour les interactions avec leurs ligands respectifs. Comme mentionné préalablement, cette ossature a été identifiée en premier dans le facteur de croissance des fibroblastes des mammifères mais aussi chez des invertébrés comme la drosophile. La résolution de la structure cristallographique des interleukines 1α et 1β a montré des similitudes de structure en β-*Trefoil*, ainsi que celle de l'inhibiteur de la trypsine de soja, des agglutinines végétales et des toxines homologues à la ricine.

ii. Protéines à motif exclusivement α

- *TPR-like*

Le motif Tétratricopeptides (Tetratrico peptide repeat : TPR) a été découvert en 1990 par Hirano et ses collaborateurs, lors de leurs travaux étudiant le gène nuc2^{+} de *S. pombe* (Hirano et *al.*, 1990). La désignation de ce motif fait référence aux 34 acides aminés qui composent la base de la répétition. Cette ossature est l'une des plus étudiées et des plus utilisées pour la génération d'interacteurs protéiques grâce à sa structure versatile (D'Andrea et Regan, 2003). C'est un motif répété qui comporte 34 acides aminés se repliant en 2 hélices α : hélice A- boucle-hélice B comme montré sur le domaine à trois motifs TPR de la

Phosphatase 5 (PP5) (Fig. 22). Ces hélices s'arrangent dans l'espace formant un angle de 24° puis s'entassent les unes contre les autres pour former une spirale droite d'hélices α antiparallèles. Chaque hélice A interagit avec l'hélice B du même *repeat* et l'hélice A' du *repeat* suivant.

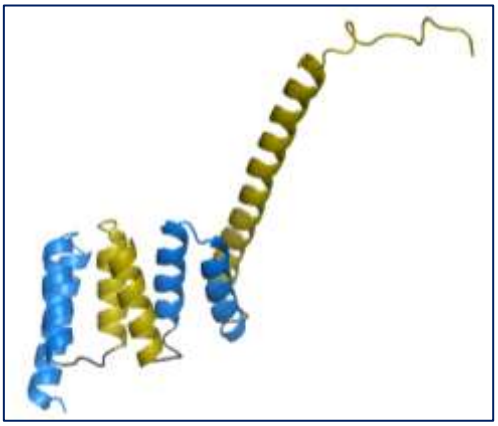


Fig. 22: Structure typique d'une protéine à motif TPR : Fragment de la phosphatase 5 humaine (PDB : 1a17).

Une séquence consensus pour ce motif a pu être déterminée à partir d'un ensemble de séquences de protéines à motifs TPR observées au sein de différents génomes (bactérie, homme, levure, plantes). Ce motif a été identifié dans différents organismes depuis les protéines bactériennes jusqu'aux protéines humaines. Ces protéines semblent être impliquées dans des processus cellulaires divers : régulation cellulaire, contrôle de la transcription, protéine du transport mitochondrial et peroxysomal, la neurogenèse et le repliement protéique.

L'étude de la distribution des nombres de répétitions de ce motif, dans les différents règnes, montre que l'organisation en 3 répétitions est la plus répandue dans la nature. Ceci reflète la nécessité d'un nombre minimum de 3 motifs pour former une protéine structurée et fonctionnelle. Toutefois, il existe des protéines contenant de 1 à 16 motifs TPR répétés. Une des caractéristiques intéressantes des motifs TPR est qu'ils ne subissent pas de changements conformationnels lors de la liaison avec le ligand. Les structures de protéines à motifs TPR présentent une sorte de rainure dotée d'une large surface d'interaction avec les ligands (Andrade, Perez-Iratxeta, and Ponting, 2001).

Ce motif a fait l'objet de travaux d'ingénierie et de construction de bibliothèques de protéines réalisés par l'équipe de Lynne Regan. Ces travaux seront décrits dans la dernière partie de cette introduction.

- *Armadillo repeat*

Le motif armadillo a été identifié pour la première fois chez la drosophile dans la protéine produit du gène de la polarité (Peifer, Berg, et Reynolds, 1994), homologue à la β-caténine humaine. Ce motif a été identifié plus tard dans différentes autres protéines telles que l'importine α, le facteur d'échange de la guanine (singGDS) (M. R. Groves et Barford, 1999). La structure cristallographique de la β-caténine à 12 motifs et l'importine α (à 10 motifs) a montré que le motif de 42 résidus est replié en 3 hélices α. Une hélice courte de 2 tours (H_1) suivie de 2 autres hélices plus longues (H_2 à 2-3 tours et H_3 à 3-4 tours). Ces dernières sont antiparallèles et s'associent les unes contre les autres, alors que H_3 leur est perpendiculaire pour former une sorte de superhélice (Fig. 23).

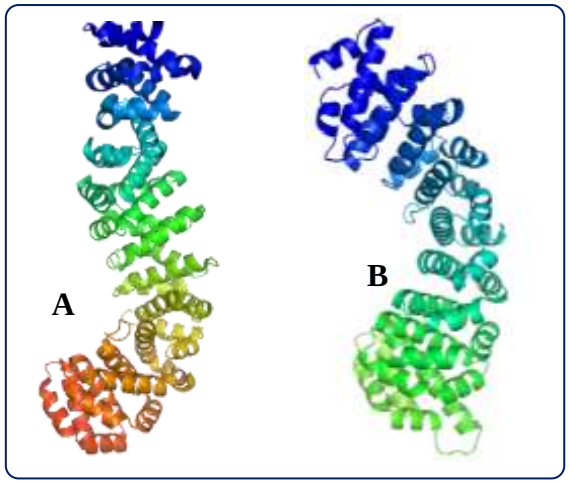


Fig. 23 : Structure de 2 protéines à motifs armadillo : A : La β-caténine humaine à 12 motifs (PDB 1JDH). B : L'α-importine à 10 motifs (PDB : 1BK5).

La fonction principale de ce motif est d'interagir avec une large gamme de protéines par le biais d'une longue fente qui s'étend tout au long de la superhélice. Cette fente est tapissée par des résidus exposés au solvant et d'autres qui sont hautement conservés. Dans l'importine α, des résidus tryptophanes sont conservés et ils sont présents au niveau du troisième tour de l'hélice H_3 des 6 motifs Armadillo. De plus 4 résidus au-delà il y a une asparagine conservées. Ces 2 résidus vont former une crête tout au long de la face concave (Andrade, Perez-Iratxeta, and Ponting, 2001).

- *Heat repeat*

Ce motif doit sa désignation aux 4 premières protéines dans lesquelles il a été identifié (Andrade, Perez-Iratxeta, et Ponting, 2001) : ***H***untingtine (impliquée dans la maladie de

Huntington), facteur d'***E***longation 3, sous unité PR65/***A*** de la phosphatase 2A, et la protéines kinase ***T***OR (target of rapamycine). Il a depuis été observé dans de nombreuses autres protéines parmi lesquelles les importines β1 et β2, le facteur d'épissage SAP155 ainsi que plusieurs familles de protéines impliquées dans des complexes associés à la clathrine, aux microtubules dans des cellules tumorales ou encore à la dynamique des chromosomes.

Sur le plan structural, c'est un motif très hétérogène dont la longueur peut varier entre 30 et 42 résidus et le nombre de répétition peut varier entre 3 et 22 (M. R. Groves et Barford, 1999). A titre d'exemple, la sous-unité PR65/*A* de la phosphatase 2A montre que cette protéine est constituée de 15 motifs Heat répétés. La protéine peut être divisée en 3 régions distinctes selon le repliement entre motifs adjacents ; région 1-3, 4-12 et 13-15. Chaque motif est composé d'une paire d'hélices α (A et B) antiparallèles de longueur équivalente, de 18 résidus chacune (fig. 24). Les paires d'hélices vont s'assembler les unes contre les autres adoptant une forme en crochet à 2 couches d'hélice. La distance entre les axes des hélices parallèles de motifs consécutifs est différente (les hélices A de deux motifs consécutifs sont plus distantes entre elles que les hélices B) ce qui induit une courbure générale de la protéine et lui confère 2 surfaces distinctes : une face concave et une convexe. La structure de chaque hélice est maintenue grâce à un réseau de liaisons hydrogènes ainsi qu'une proline conservée sur l'hélice A (Andrade et *al.*, 2001). L'empilement des motifs adjacents se fait par le biais de liaisons hydrophobes ce qui définit une sorte de corps hydrophobe continu tout au long du solénoïde. Comme mentionné au préalable, ce motif est très hétérogène, ainsi aucune étude n'a abouti à la conception d'un consensus ni la construction de banque de protéines artificielles basées sur ce motif.

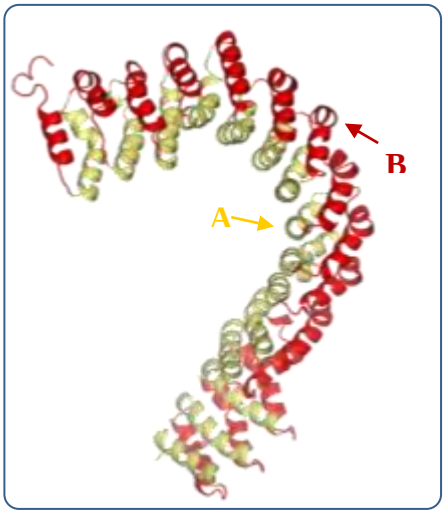


Fig. 24 : Structure de la sous unité PR65/A de la phosphatase 2A (PDB : 1b3u) typique d'une protéine à motifs *HEAT repeat* : Les 2 hélices A et B constituant le motif sont représenté respectivement en jaune et rouge.

iii. Motifs mixtes α-β

- *Leucines Rich Repeats (LRR)*

C'est un motif qui est relativement court par rapport aux motifs précédemment décrits. La longueur du motif LLR varie entre 20 et 29 résidus (Kobe et Deisenhofer, 1995). Il a été découvert chez l'α2-glycoproteine riche en leucine à partir du sérum humain. Cette protéine est composée de 312 acides aminés dont 66 sont des leucines. Sa séquence peut être divisée en 13 fragments de 24 résidus chacun (Takahashi, Takahashi, and Putnam, 1985). De manière générale, ce motif est répété en tandem au sein d'une protéine et le plus grand nombre de répétition décrit est 30 répétitions, dans la protéine choaptine chez la drosophile. La structure tridimensionnelle de ce motif a été décrite à partir de la structure résolue de l'inhibiteur de la ribonucléase à 15 repeats. Un seul motif consiste en un court brin β et une hélice α (Fig. 25). Les motifs se succèdent les uns après les autres, parallèlement à un axe commun, adoptant une forme courbée rappelant un fer de cheval. Cette structure est maintenue grâce à l'empilement des résidus consensus et la formation du corps hydrophobe de la protéine.

Les hélices, dans cette protéine, forment une circonférence externe (face convexe) alors que les brins s'organisent en une sorte de feuillets β parallèles tout au long de la circonférence interne (face concave) de la protéine formant une surface d'interaction potentielle. Il faut, toutefois, signaler que cette face est le point commun entre les LRR, dont les variétés peuvent être très hétérogènes. En effet, la face convexe de la protéine peut être formée par l'empilement d'hélices α (de différentes longueurs) comme chez l'inhibiteur de la ribonucléase ou encore par des régions entre-brins adoptant une sorte de boucle comme observé chez la protéine YopM (Kobe and Kajava, 2001).

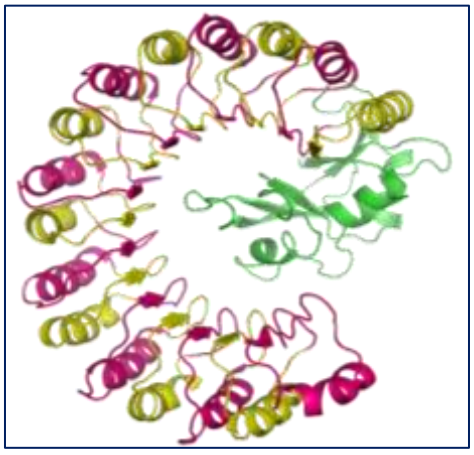


Fig. 25 : Structure du complexe de la ribonucléase (en vert) avec son inhibiteur (PDB : 1dfj) : l'inhibiteur de la ribonucléase porcine présente une structure typique d'une protéine à motif LRR.

Ce motif est exclusivement impliqué dans des interactions protéines-protéines. Les protéines à motifs LRR sont observées en transduction de signal, adhésion cellulaire, réparation d'ADN, recombinaison, transcription et maturation d'ARN : depuis la bactérie jusqu'à l'homme. Ce motif a fait l'objet des travaux très récents de Lee et ses collaborateurs pour la conception d'une nouvelle ossature : « *Repebody* » que nous développerons dans la partie suivante.

- *Ankyrines*

Le motif ankyrine doit sa désignation à la première protéine dans laquelle il a été décrit : l'ankyrine érythrocytaire humaine (Andrade, Perez-Iratxeta, et Ponting, 2001). Ce motif, comprenant 33 résidus, a été décrit à partir de la première structure résolue d'une molécule en motifs ankyrines : le complexe 53BP2 avec p53. Il est organisé en une structure en épingle à cheveux β-hélice α-une boucle-une hélice (β2α2) (Sedgwick et Smerdon, 1999). Sur le plan structural, les résidus hydrophobes du corps des hélices α forment des surfaces non polaires s'étendant en un faisceau hélicoïdal stable (Fig. 26). De plus, les liaisons hydrogènes établies entre les motifs adjacents renforcent la structure de la protéine. Les chaînes latérales de petite taille, revêtant la face intérieure et l'épaulement gauche des hélices empilées, vont quant à eux constituer une surface caractéristique accessible au solvant.

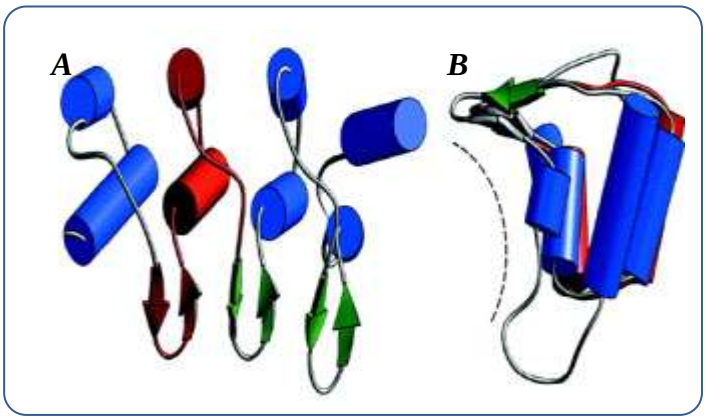


Fig. 26 : Structure de la 53BP2 typique d'une protéine à motif ankyrine : Le motif se replie en 2 hélices α (cylindres bleus et rouges) et une boucle β (flèches vertes). A est une vue de dessus et B vue du côté montrant la poche (arc discontinu) que forme l'empilement des motifs en particulier les boucles.

Ce motif est ubiquitaire, il est impliqué essentiellement dans l'interaction avec d'autres protéines appartenant à une large gamme de classes protéiques comme par exemple le suppresseur nucléaire de tumeurs p53, la protéine kinase impliquée dans la division cellulaire CDK6 ainsi que la sous-unité non catalytique M130 et l'unité catalytique PP1c de la myosine phosphatase des muscles lisses. En effet la résolution des structures de plusieurs

complexes a révélé que les protéines à motif ankyrine semblent interagir par leur fente formée par l'empilement des feuillets β des motifs adjacents.

Conclusion

Les 6 familles de protéines à motifs répétés, décrites ci-dessus, sont caractérisées par leur capacité de former des structures monomériques et non fibreuses. Toutefois, d'autres familles de motifs répétés adoptent des structures plutôt fibreuses. On peut citer à titre d'exemple, les domaines à motifs répétés des protéines de staphylocoque fixant la fibronectine. Ces motifs non structurés en solution n'adoptent des structures tertiaires bien définies qu'en présence de leur ligand ; la fibronectine extracellulaire des mammifères. D'autres structures rigides et oligomériques existent aussi bien en superhélice (2 à 5 hélices amphiphiles) ou encore en long filaments de structures β telles les protéines des fibres des adénovirus.

c) Les motifs répétés et la conception de protéines artificielles :

Les ossatures protéiques basées sur les motifs **ankyrine**, **TPR**, **LRR**, Armadillo et dans ce travail le motif **Heat**, ont été exploitées pour la création de banques de protéines artificielles. Les banques d'ankyrines et, plus récemment, des LRR, ont été utilisées pour l'identification de variants reconnaissant spécifiquement une large gamme de molécules cibles. Pour l'étape de conception des banques basées sur ce type de protéines artificielles la détermination d'un consensus pour le motif, qui se trouve répété, est une étape cruciale. Un consensus, bien conçu, récapitule les éléments de séquence essentiels à la conformation de chaque module et aux interactions des modules entre eux permettant ainsi le maintien de la stabilité intrinsèque des protéines de la banque. Les motifs basés sur un même consensus sont compatibles entre eux ce qui autorise l'ajout, la délétion et les échanges de motifs au sien de la banque. Et finalement, concevoir « des protéines consensus », dans lesquelles les motifs ont la même séquence, permet d'étudier les principales caractéristiques de l'ossature protéique ce qui facilite les étapes ultérieures d'ingénierie.

La détermination d'une séquence consensus se fait par le choix de l'acide aminé le plus fréquent pour une position donnée dans l'alignement des séquences d'un même groupe de « *repeats* » disponibles dans les bases de données. L'hypothèse de départ pour la définition d'un consensus est que les résidus stabilisants sont conservés. L'approche consensus n'est pas

réservée aux protéines à motifs répétés dès lors qu'un nombre suffisant de séquences est disponible. La première réussite de cette approche a été obtenue par les travaux de Berg sur le design d'un consensus de peptide en doigt de zinc. Plus tard, l'ingénierie de consensus plus compliqués a été appliquée pour des domaines variables d'IgG ou des domaines SH3 dans le but d'augmenter leur stabilité (Kajander, Cortajarena, et Regan, 2006). Et en étape ultime ce concept a été appliqué aux protéines à motif répétés telles que les ankyrines, les TPR, LRR et au cours de notre travail les *Heat repeats*.

L'élaboration d'un consensus passe essentiellement par une étape d'extraction d'un ensemble non redondant de séquences du *repeat* étudié puis par la génération d'un multialignement de ces séquences. Cette étape cruciale peut être réalisée manuellement ou encore de façon automatique grâce aux alignements préexistants dans les bases de données (SMART, PFAM). L'alignement manuel est requis par exemple pour séparer l'alignement des motifs internes, de celui des « *cap* » ou encore lorsque la définition d'une famille repose sur une définition très large incorporant des sous-groupes de séquences non compatibles entre elles.

Une fois l'alignement de séquences réalisé, une étude statistique des distributions des types d'acides aminés par position peut être réalisée par une simple feuille de calcul Excel ou par divers outils disponibles sur le web. Ainsi on peut identifier à la fois la variabilité, la fréquence et la nature des résidus à chaque position. Et par la suite, on peut choisir soit d'exploiter le consensus ainsi défini : retenir les résidus les plus conservés pour chaque position. Ou encore de retenir l'acide aminé le plus conservé pour les positions conservées alors que les positions variables dans l'alignement de séquences seront l'objet de différentes stratégies de mutations.

Dans la définition des résidus du consensus, on peut soit prendre en considération la fréquence des acides aminés à une position données ou encore la proportionnalité globale qui permet de prendre en compte les acides aminés les plus rares (Main et *al.*, 2003). Cette dernière approche, a été par exemple, adoptée pour la conception des consensus des TPR. Un calcul de la propension (P_g) a été effectué pour chaque position : P_g représente le ratio entre la fréquence d'un acide aminé à une position et sa fréquence dans toute la protéine. Ce calcul va permettre de déterminer la préférence naturelle d'un acide aminé à une position donnée par rapport aux autres positions de la séquence. Cette équipe a suivi les 2 stratégies pour concevoir et construire des domaines TPR à 3 motifs. Les deux approches ont donné des

domaines qui sont plus thermostables que les TPR naturels qui adoptent la structure typique des TPR.

Lors de la définition d'un consensus par les méthodes décrites ci-dessus chaque résidu est considéré de façon isolée. Toutefois, la relation entre la conservation des acides aminés dans les séquences et la stabilité des protéines obtenues n'est pas aussi évidente : dans une structure repliée, chaque chaine latérale peut interagir avec les résidus voisins. Les effets de la covariance peuvent être significatifs (Magliery and Regan, 2004). En effet, l'existence d'un acide aminé à une position donnée peut influencer l'existence d'un autre acide aminé plus loin dans la séquence si celui-ci est voisin dans l'espace, soit pour former des interactions favorisant le repliement et la stabilité de la protéine soit pour éviter les encombrements stériques déstabilisants. La prise en compte des covariances revient à calculer la fréquence observée d'une paire de résidus définis, par exemple des positions voisines dans la structure, relativement au produit des fréquences de ces mêmes résidus, considérés isolément. Il est concevable ainsi de parvenir non pas à un consensus sous forme d'une séquence unique mais à une définition plus large d'une famille de séquences qui obéiraient néanmoins à des règles structurales communes. L'identification exhaustive des covariances suppose cependant que le nombre de séquences non redondantes d'une même famille structurale soit assez grand pour échantillonner de façon significative toutes les paires de résidus possibles à deux positions quelconques, ce qui dans les fait est assez rarement rencontré.

La longueur de la protéine, reliée au nombre de motifs répétés, doit également être considérée. Ainsi, Regan et ses collaborateurs ont fait le choix de produire des protéines à 3 motifs TPR. En effet, c'est la longueur la plus fréquente pour les domaines TPR et la plus petite pour les unités fonctionnelles naturelles. Dans la même logique, Plückthun a plutôt choisi de créer des protéines à 5 et 6 motifs ankyrines : 2 à 3 motifs internes en plus du *Ncap* et du *Ccap*.

Des structures cristallographiques résolues de TPR naturelles ont montré une hélice en plus du côté C-terminal du dernier motif répété. Ainsi Regan et ses collaborateurs ont suggéré l'importance d'une hélice terminale appelée « *cap* », jouant un rôle important dans la stabilité, le repliement et la solubilité de la protéine à motifs répétés (Main et *al.*, 2003). De la même façon, il a été constaté que la solubilité des ankyrines est fortement liée à la présence des motifs caps des deux côtés N et C-terminaux qui sont inspirés de modèles naturels. En effet, des travaux sur des ankyrines sans « *cap* » ont montré que celles-ci sont insolubles et

s'expriment en corps d'inclusion (Mosavi, Minor, et Peng, 2002). Les motifs d'extrémités ou «*cap*» semblent être alors des éléments essentiels à prendre en compte lors la conception de protéines à motifs répétés.

Une étude très intéressante a été réalisée par l'équipe de Regan, concernant l'analyse statistique de la distribution des acides aminés dans les domaines TPR. En effet, cette famille se prête bien à une étude statistique de par le nombre important de séquences disponibles dans la base de données (autour de 10000 dans la PFAM). Cette équipe a eu recours au calcul de l'entropie relative, reliée à la variabilité par position, puis a reporté ce paramètre sur la structure expérimentale de complexes de domaines TPR et leurs ligands. Cette présentation montre que la face concave du domaine est plus variable que la face convexe plus exposée au solvant. Pour les deux domaines TPR pris comme exemples, les résidus en contact avec le ligand correspondent aux positions de faible entropie témoignant d'un écart négligeable par rapport à l'état de référence. Ces positions particulières comportent donc une part réduite de l'information spécifiant l'architecture du module répété, ce qui autorise une variabilité adaptée aux interactions spécifiques contractées par chaque protéine. Sur une échelle plus grande, une analyse statistique d'un plus grand nombre de motifs TPR confirme que les positions impliquées dans l'interaction avec les ligands sont substituées aléatoirement par les acides aminés favorables à l'interaction. Parmi les positions variables impliqués dans la reconnaissance des peptides ligands (par exemple les postions : 2, 5, 6, 9, 12, 13 sont impliquées dans l'interaction de HOP TPR1 et TPR2A avec leur ligands (Fig. 27) certaines positions sont plus biaisées que d'autres (Magliery et Regan, 2005). La position 6, en particulier, est plus conservée que les autres positions. Correspondant le plus fréquemment à une Asn, cette position est plus enfouie dans la structure que les autres positions nécessaires à l'interaction et elle est utilisée pour établir des liaisons hydrogène avec le squelette du ligand peptidique.

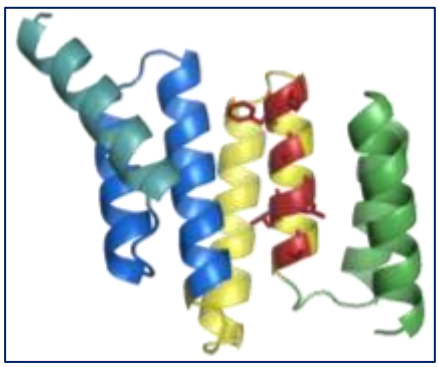


Fig. 27 : Structure de la TPR2A montrant les acides aminés impliqués dans l'interaction avec son ligand.

Afin de valider l'approche, cette équipe l'a appliquée à une autre famille de protéines à motifs répétés : les ankyrines. En effet, l'analyse de différents domaines ankyrines a permis de confirmer que la détermination des entropies relatives basses permet d'identifier les résidus responsables de la spécificité Protéine-ligand. En effet, les résidus d'interaction avec les ligands sont situés sur la boucle β et la région proche : les positions 2, 3, 5, 13, 14, 17, 32, et 33 (Fig. 28).

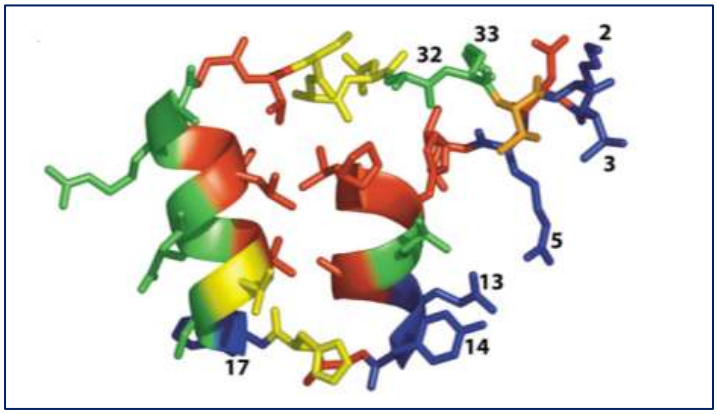


Fig. 28 : Site d'interaction d'un motif ankyrine montrant les résidus impliqués dans l'interaction (Extrait de la structure du domaine Ank de la GABPβ1 avec son ligand GABPα (Magliery and Regan, 2005)).

La conception du consensus pour cette famille est passée par plusieurs étapes (Mosavi, Minor, and Peng, 2002). L'identification des positions du consensus ankyrine conservées et moins conservées est basée sur l'étude de Sedgwick et Smerdon (Sedgwick and Smerdon, 1999). Cette étude faite à partir de données accumulées grâce aux structures résolues de complexes avec des motifs ankyrines (3000 motifs Ank à partir de 400 protéines non redondantes), montre que certains résidus conservés sont impliqués dans la structuration et la stabilisation des motifs comme entité à part entière alors que d'autres permettent l'encrage des motifs les uns avec les autres. La structure en épingle à cheveux est stabilisée par le biais des ponts H établie entre l'aspartate de la position 1 et le résidu canonique de la position 33. Ce consensus contient 3 Glycines hautement conservées ; une à la position 4 terminant la boucle β2 et 2 aux positions 15 et 27 terminant les 2 hélices de ce motif. On note aussi l'existence d'une séquence caractéristique (Thr-Pro-Leu-His) qui forme une sorte d'épaulement au début de la première hélice. L'empilement des motifs les uns aux autres va orienter les structures β en épingle à cheveux de sorte que la β2 d'un motif interagit avec la β1 du motif suivant. A partir de ces observations l'équipe de Plückthun, ont aligné 229 motifs Ank de 33 acides aminés et sans ajout ni délétion, ce qui a donné le 1er consensus A contenant les résidus 3 à 32. L'étape suivante visait à définir les résidus aux positions manquantes et celles qui sont sans préférence définie (à fréquence < à 30%), ce qui a été réalisé à partir d'alignement de

motifs Ank de structure résolue pour donner le consensus B. Ce dernier consensus a été soumis à un BLAST dans GenBank. Les 200 séquences homologues les plus proches ont été alignées et analysées manuellement pour donner le consensus C. La fréquence des résidus par position a été confirmée par la suite avec un alignement de 2220 séquences d'Ank de la base de données PFAM. L'ultime étape d'affinement du consensus s'est faite par analyse de bases de données structurales comprenant 10 structures résolues d'ankyrines permettant ainsi la définition des résidus pouvant servir de support aux interactions avec les partenaires.

L'analyse, par le programme NACCESS, des changements survenus sur la surface accessible au solvant de chaque résidu lors de la formation des complexes de structure résolue, a permis de définir les positions d'interaction avec les cibles. Ainsi, on observe que le coude β et la première hélice sont les régions impliquées et plus précisément les positions 2, 3, 5, 13, 14 et 33. Dans la bibliothèque réalisée, ces positions sont codées pour pouvoir conduire à n'importe quels résidus sauf la Glycine et la Proline pour des raisons d'ordre structural. La Cystéine est également absente du codage pour éviter la formation de pont S-S. Les 27 positions restantes formant le motif Ank sont alors définies. Les positions 1 et 4 sont respectivement Asp et Gly (selon la fréquence de ces acides aminés). La séquence caractéristique (TPLHL) entre les positions 6 et 10, est maintenue dans le consensus. Ceci revient au fait que la His9 établit des liaisons hydrogène stabilisantes avec Thr6, l'Ala32 et le résidu de la postions variable 5 du motif suivant. La Pro7 coupe l'hélice1 pour s'intercaler dans le corps hydrophobe. Et finalement la Leu8 se lie au corps hydrophobe en pointant vers l'hélice 2 et le motif suivant. Les positions 11 et 12 sont des Ala nécessaires pour le maintien de la structure générale du motif. La position 15 est une Gly pointant à partir de l'hélice 2. L'His16 est partiellement enfouie et forme des liaisons H latérales avec l'hélice 1 du motif précédent et d'autres liaisons avec les chaînes principales de l'Ala11, l'Ile19 et la Val20. La position 17 est définie comme Leu car c'est la plus fréquente (27%) et pour des considérations structurales en tant qu'initiateur de l'hélice2. Cette dernière est amphiphile et elle ne peut être bien définie à partir de la base de données, Plückthun et ses collaborateurs ont alors eu recours à d'autres paramètres. A titre d'exemple, la Glu18 a été choisie car elle apparait régulièrement dans l'inhibiteur cdk4 (p18) et qu'elle est tolérée structuralement. Les leucines des positions 22, 23, 24 sont une entité conservée de la partie supérieure du corps hydrophobe du motif. On peut citer aussi la position 26 est très ambigüe : Asn est la plus abondante, Ala est mise en évidence à partir des alignements de séquence, His et Tyr peuvent aussi être une alternative puisque c'est une position clé pour maintenir la distance entre les motifs et par la suite ce

résidus doit laisser assez d'espace et être probablement un acide aminé polaire. L'équipe de Plückthun a alors opté pour une combinaison His-Tyr-Asn, ce qui est facilité par le fait que ces acides aminés peuvent être codés par le même codon HAC. Ainsi une version finale du consensus a été établie : consensus D. Cette équipe s'est ensuite intéressée à la conception des motifs « *cap* ».

En effet, ces motifs de coiffe (*cap repeats*) permettent de masquer les deux extrémités du cœur hydrophobe. Ces motifs *Ncap* et *Ccap* peuvent s'intégrer dans l'empilement des motifs de la protéine, par une de leur face qui comporte des résidus hydrophobes nécessaires. Toutefois, ils diffèrent des motifs internes par leur autre face dirigée vers l'extérieur, « l'exoface », qui est composée de chaînes latérales polaires qui peuvent demeurer exposées au solvant. L'équipe a retenu des motifs cap naturels interagissant avec des motifs proches du consensus garantissant ainsi une structure connue et une compatibilité probable avec les autres modules. Ainsi pour le *Ncap* est une adaptation du motif GABPβ1 de souris alors que *Ccap* est adapté à partir du domaine de 1AWC_B.

d) De la conception théorique à la synthèse moléculaire :

Plusieurs approches ont été proposées dans le but de créer des bibliothèques de protéines basées sur le consensus de protéines à motifs répétés. Les approches utilisées pour construire les protéines à motifs répétés exploitées expérimentalement **TPR** (Regan), **ankyrines** (Plückthun) et les **LRR** (Kim H-S) sont décrites dans la section suivante.

i. Les banques de TPR

Pour les banques de protéines artificielles à motifs TPR, L. Regan et ses collaborateurs ont opté pour des protéines de longueurs préalablement fixées contenant 1, 2 ou 3 motifs. Dans la nature, les domaines à 3 motifs répétés sont les plus petits et les plus répandus. Les protéines à 1 et 2 motifs sont construites par une amplification de 4 oligonucléotides à extrémités chevauchantes suivie par une PCR avec deux primers additionnels correspondant aux séquences terminales. Une approche différente a été suivie pour les protéines à 3 motifs. En effet, elles sont produites par une simple ligation de 3 cassettes d'ADN double-brin, correspondant chacune à une unité TPR, le produit de ligation est par la suite amplifié par une PCR. Ces gènes sont construits avec des sites de restriction permettant de les insérer dans le vecteur d'expression. Ce dernier contient une étiquette His du côté N-terminal suivie d'un site

de clivage à la TEV protéase permettant de l'éliminer après la première étape de purification par affinité (Main et *al.*, 2003).

Ces protéines à motifs TPR ont été très étudiées par l'équipe de Regan et en particulier le lien séquence-stabilité thermique telle qu'elle peut être mesurée par des techniques de microcalorimétrie. Des protéines contenant un nombre croissant du même motif TPR : de 2 à 20 ont été construites et leurs thermogrammes de dénaturation ont été déterminés par DSC. Ces derniers ont montré les caractéristiques essentielles suivantes:

- plus une protéine contient d'insert, plus la température de demi-transition est élevée, donc plus elle est stable.
- la surface sous le pic endothermique est d'autant plus grande que la protéine contient plus de motifs insérés.

La réversibilité de la dénaturation a aussi été vérifiée. En effet, pour toutes les protéines TPR, deux séries successives de dénaturation thermique ont été réalisées. Leur superposition parfaite démontre la réversibilité de la dénaturation. Les données de microcalorimétrie ont permis la compréhension du mécanisme de dénaturation. Le modèle simple en 2 états pour l'ajustement des thermogrammes correspond parfaitement pour les protéines de 1 et 2 motifs mais au-delà de 3 motifs ce modèle théorique n'est plus adapté. Ceci plaide en faveur de l'hypothèse de l'existence d'états intermédiaires au cours de la dénaturation. De plus, l'asymétrie des thermogrammes devient de plus en plus prononcée avec l'augmentation du nombre de motifs insérés dans la protéine TPR, ce qui peut être également expliqué par la présence d'espèces intermédiaires. Enfin, la capacité calorifique observée en DSC augmente linéairement avec le nombre d'acides aminés par protéine. Par comparaison avec les protéines globulaires, ces TPR ont une ΔCp plus petite ce qui indique que la surface exposée, lors de la dénaturation, est moindre que la surface estimée à partir des protéines globulaires (Cortajarena et Regan, 2011).

L'équipe de Regan a également construits une banque de variants TPR, comportant $2.7*10^8$ clones puis a sélectionné des variants interagissant spécifiquement des cibles préalablement choisies. Cette équipe a proposé une stratégie de criblage reposant sur la reconstitution de la GFP fluorescente. En effet, dans ce système les deux fragments complémentaires de GFP sont fusionnés génétiquement l'un à la banque de TPR et l'autre à la

protéine cible. Seuls les clones où la paire TPR-cible interagit effectivement donnent lieu à la reconstitution de la GFP (*split-GFP reassembly assay*). Ces clones peuvent être alors identifiés et triés par fluorescence en cytométrie de flux (Jackrel et *al.*, 2009).

En premier lieu, des sélections ont été réalisées en utilisant comme cible le tag cMyc ainsi que la protéine Dss1 (connue pour interagir avec le suppresseur de tumeurs BRCA2). Ces sélections ont révélé un grand nombre de variants TPR qui reconnaissent Dss1. Ils ont par la suite effectué une analyse de séquences permettant de comprendre le mécanisme de l'interaction et les résidus mis en œuvre, en particulier ceux qui sont hypervariables. Pour étudier plus en détail les interactions, 2 clones ont été choisis et ils ont révélé des constantes de dissociation de l'ordre de 10 µM. Par une étude biophysique (CD et microcalorimétrie), il a été vérifié que ces clones gardent les caractéristiques structurales attendues pour les protéines de cette famille et qu'il n'y a pas de relation entre l'affinité d'interaction et la stabilité de la protéine ce qui indique que la variation de séquence dans les positions hypervariables n'affecte pas la structure générale de la protéine.

Par la suite, cette équipe a proposé une démarche de maturation d'affinité : en partant d'un clone de la banque qui interagit avec une cible définie, des cycles de sont réalisés par *error prone PCR* dans le but d'introduire des mutations dont certaines pourraient améliorer l'interaction initiale. Une amélioration a effectivement été obtenue : la constante de dissociation K_D d'un variant initialement de 180 µM a pu être diminuée à 20 µM.

ii. Les banques ankyrines :

Pour la réalisation des banques d'ankyrines, le groupe de Plückthun a eu recours à une stratégie différente de celle suivie par Regan pour les TPR. En effet les positions constantes du module ankyrine consensus ainsi que les *Ncap* et *Ccap* ont été codés par des codons optimisés pour l'expression chez *E.coli.* Sur les oligonucléotides utilisés, les codons correspondant aux positions hypervariables des motifs ankyrines (2, 3, 5, 13, 14 et 33) ont été synthétisés par assemblage de précurseurs de trinucléotides, plutôt que de mononucléotides comme c'est habituellement le cas. Pour une position dégénérée est alors simplement ajoutée un mélange, en une proportion choisie de 20 trinucléotides, correspondant aux 20 acides aminés (Virnekäs et *al.*, 1994). Si l'on souhaite maintenir une variabilité partielle, par exemple pour omettre les résidus Pro ou Cys, il est possible de n'ajouter que 19 précurseurs de trinucléotides. La même opération est plus délicate à réaliser sans cette technologie de

synthèse. Par cette méthode, les positions dégénérées ont été alors codées en A, D, E, H, K, N, Q, R, S, T avec une probabilité de 7% chacun et F, I, L, M, V, W et Y à 4% chacun. Le but principal était d'exclure les acides aminés incompatibles avec la stabilité des motifs (Gly et Pro) d'éliminer les cystéines et d'éviter une hydrophobie moyenne trop élevée aux positions de surface. Enfin, la position 26 a été définie pour le codon HAC pour His, Tyr et Asn. L'assemblage des motifs s'est fait par des étapes de PCR à partir des oligonucléotides conçus suivant la stratégie consensus. Le *Ncap* et le *Ccap* ont été assemblés séparément avec les nucléotides qui leur correspondent. Le *Ccap* est ligué dans le plasmide d'expression dérivé du *pQE30*. La casette *Ncap*, non insérée dans un vecteur, est liguée à un module codant un seul motif ankyrine pour permettre de produire une séquence N1 où le N désigne le cap N-terminal et 1 le nombre de motif ankyrine. Le produit de ligation (*Ncap* motif 1) est par la suite amplifié par PCR par des oligonucléotides spécifiques. Il est alors clivé spécifiquement à la fin du motif1 puis ligué avec un second module ankyrine pour entamer un nouveau cycle d'amplification. Ces étapes sont répétées jusqu'à l'obtention du nombre de motifs désiré (souvent 2 ou 3) avec le *Ncap*. Pour finaliser la construction des protéines ankyrines, les modules assemblés avec le *Ncap* sont insérés dans le vecteur comportant le *Ccap* et reconstituent ainsi la séquence codante entière. L'ensemble de la construction utilise des sites de restriction de type 2S (comme BsaI) générant des extrémités cohésives non symétriques, ce qui permet l'assemblage des cassettes dans une seule direction. Un calcul simple permettrait d'estimer la diversité théorique de la banque : on a 7 positions hypervariables par motif ce qui correspond à $(3*17^6) = 7.2\ 10^7$ variants /motif et donc pour les ankyrines de 2 motifs $(3*17^6)^2 = 5.2*10^{15}$ variants et pour celles à 3 motifs $(3*17^6)^3 = 3.8*10^{23}$ variants. La combinatoire des motifs génère donc une diversité très élevée, sans compromettre davantage la stabilité de l'édifice. Ainsi des banques combinatoires de protéines artificielles à motifs ankyrines répétés ont été construites. L'équipe de Plückthun désigne ces banques de protéines artificielles par DARPins : *Designing Ankyrin Repeat Proteins*.

Comme étape suivante, cette équipe a effectué une étude complète dans le but de valider la stratégie de conception et de construction moléculaire de ces banques de protéines artificielles. Ainsi les séquences des clones pris au hasard des banques ont été analysées. 18 clones seulement parmi 52, ont une séquence d'ADN correcte avec une proportion d'erreur plus importante lorsque la protéine contient plus de motifs insérés. Les erreurs sont localisées au niveau du *N-cap* et elles sont dues probablement à la qualité des oligonucléotides. Les erreurs localisées au niveau des motifs ankyrines insérés sont des décalages du cadre de

lecture. Une analyse des séquences et plus spécifiquement des positions hypervariables a été réalisée. Ces positions ont une distribution approximativement aléatoire et aucun codon non désiré n'a été détecté (Gly, Pro Cys, stop). 75% des motifs ankyrines séquencés ont des séquences correctes.

L'étape ultime de cette caractérisation est la détermination des propriétés biophysiques de ces protéines artificielles. Des clones codants de séquences connues et de différentes tailles ont été utilisés pour vérifier l'expression et la solubilité ainsi que la structure secondaire et la stabilité thermique. Ces clones s'expriment sous forme soluble à un taux important. Ils ont été purifiés en une seule étape par chromatographie d'affinité et leurs masses ont été déterminées par *SDS page* et confirmées par MS. L'état d'oligomérisation des protéines a été déterminé par tamis moléculaire : 5/6 sont monomériques alors que la 6^{éme} formée un équilibre entre monomère et un dimère dans la solution. Une de ces ankyrines artificielles contenant 3 inserts a vu son état monomère confirmé suite à la résolution de sa structure. Suite à la détermination des spectres CD, on a vérifié que la structure secondaire des protéines correspondait à ce qui a été attendu. La stabilité thermique des ankyrines a été étudiée : ces protéines ont une dénaturation coopérative, partiellement réversible et ont des températures de demi-fusion élevées (entre 66 et 85°C) (Binz et *al.*, 2003). Ces protéines sont donc solubles, repliées, très stables et ont été utilisées par la suite pour trouver des variants capables de reconnaître spécifiquement et avec une haute affinité des molécules cibles préalablement choisies.

Dans cette perspective, l'équipe a réalisé une série de travaux remarquables sur cette famille de protéine (les *DARPins*) et qui démontrent l'intérêt de cette approche.

Les premières sélections d'interacteurs spécifiques à partir des banques d'ankyrines artificielles ont été réalisées par *ribosome display* sur des cibles modèles comme la *Maltose binding protein*. Il est apparu que des interacteurs spécifiques ayant des affinités de l'ordre du nM pouvaient être obtenus sur cette cible modèle, puis sur un ensemble très divers d'autres cibles d'intérêt. Les sélections par *phage display* sont également possibles et donnent efficacement des interacteurs d'affinités comparables à celles des interacteurs obtenus par *ribosome display*. Une condition essentielle aux sélections par *phage display* est l'utilisation de la séquence d'export de type SRP, ce qui permet l'exportation périplasmique des *DARPins* dont la stabilité conformationelle est très élevée.

Les sélections par *ribosome display* permettent des améliorations ponctuelles liées à des mutations situées hors des positions hypervariables. En effet lors de l'étape d'amplification, il est possible de récréer par *error-prone PCR* des variations qui peuvent ultérieurement être sélectionnées pour réaliser un processus de la maturation de l'affinité et de la spécificité et atteindre des affinités picomolaires.

Les applications possibles de ces protéines sont variées et concernent des domaines aussi bien en thérapeutique qu'en diagnostic. Ces protéines sont également potentiellement intéressantes comme chaperons de cristallisation, pour favoriser les études structurales y compris pour la cristallogenèse de protéines membranaires. Des réalisations importantes ont déjà été menées dans plusieurs de ces domaines. A titre d'exemple, plusieurs *DARPins* ont été fusionnées à l'hémagglutinine dans le but de rediriger des vecteurs lentiviraux vers les cellules tumorales pour détruire (Munch et *al.*, 2011). Les *DARPins* ont été utilisées aussi pour le diagnostic par immunohistochimie de tumeurs : entre autres, une *DRPin* spécifique de HER2, ayant une affinité de l'ordre du picomolaire, s'est avérée capable de détecter une amplification de l'expression de HER2 avec une sensibilité similaire à celle d'un anticorps de routine mais avec une plus grande spécificité (Theurillat et *al.*, 2010).

Au niveau thérapeutique, une *DARPin* spécifique de la molécule d'adhésion de cellules épithéliales (EpCAM) a été utilisée en tant que transporteur d'un Si-ARN complémentaire à la *bcl-2* ARNm, un facteur pro-aptotique. Plusieurs stratégies moléculaires ont été adoptées et elles ont toutes permis la diminution de l'expression de *bcl2* entrainant ainsi une sensibilisation importante des cellules MCF-7 $EpCAM^+$ à la doxorubicine. Cet effet, n'est pas obtenu dans les cellules qui n'expriment pas la protéine cible ($HEK293TEpCAM^-$) (Winkler et *al.*, 2009).

iii. Les banques de LRR ou « Repebodies »

Le groupe de Kim (S.-C. Lee et *al.*, 2012) a proposé, pour la construction de leur banque de protéines artificielles basées sur la répétition d'un motif LRR et inspirées à partir des récepteurs lymphocytaires variables des lamproies (VLR : Variable Lymphocyte Receptors), une stratégie toute à fait différente de celles de Regan et Pluckthun pour leurs banques respectives. Comme décrit au préalable (3.6.), les VLR sont constitués :

$$LRRNT\text{-}LRR1\text{-}[LRRV]_n\text{-}LRRVe\text{-}CP\text{-}LRRCT$$

Ils ont tout d'abord conçu une ossature modèle, en suivant ce schéma général. La séquence consensus des LRRV, a été déterminée par des alignements de séquence de deux bases de données (1.000 LRR de UnitProt database et 439 motifs de NCBI) :

LTNLTXLXLXXNQLQSLPXGVFDK

Dans cette séquence, les positions en noir (10 positions) sont hautement conservées, celles (3 positions) représentées en orange ont été fixées à N, Q et K puis que ces résidus sont rencontrés dans une large gamme de séquences des deux bases de données. Alors que les 6 positions représentées en bleu sont les moins conservées et elles ont été désignées de telle sorte que les acides aminés les plus fréquents et les moins chargés soient représentés. Les X sont les positions hypervariables et ils ont été fixés aux acides aminés les plus fréquents pour la construction de l'ossature modèle. Le nombre de répétition des LRRV a été fixé à 5 répétitions pour cette ossature puisque c'est le nombre le plus représenté dans les VLR naturels. Le gène correspondant a été synthétisé avec les codons optimisés pour l'expression chez *E.coli* : une protéine de 29 kDa exprimée soluble à 2 mg/L de culture. Cette ossature modèle a alors permis de valider la stratégie générale, puisqu'elle s'exprime sous forme soluble mais à faible taux.

L'étape suivante été alors de parvenir à atteindre un taux d'expression important, ils ont alors opté à l'amélioration du domaine N-terminal du *scaffold*. Par des approches rationnelles, ils ont alors décidé d'utiliser le *Ncap* de l'internaline B. Suite à des optimisations computationnelles, une deuxième ossature a été obtenue : *Repebody* (module répétés basés sur des anticorps). Cette dernière est composée de :

$[LRRNT\text{-}LRR1\text{-}LRRV1]_{internalineB}$ -$[LRRV]_4$-LRRVe-CP-LRRCT.

Cette démarche a permis d'atteindre, un taux d'expression soluble de 60 mg par litre de culture. Ils ont aussi vérifié la stabilité, thermique et en fonction du pH de leur protéine en particulier en fonction du nombre de motifs LRRV insérés. La stabilité thermique de cette ossature de deuxième génération semble augmenter avec le nombre de motifs insérés (entre 3 et 6). La température de demi-fusion pour la protéine à 6 motifs est de 85°C et ces « *repebody* » semblent stables pour une large gamme de pH entre 3 et 12.

Comme étape suivante, l'équipe s'est fixée comme objectif de vérifier la possibilité de générer des interacteurs à partir de cette ossature particulière. Ils ont alors effectué un transfert de sites d'interaction connus pour deux cibles choisies, sur cette ossature (en particulier au niveau des 4 motifs LRRV répétés) à partir de VLR naturels. Cette approche leur a permis d'obtenir 2 « *repebodies* » artificiels reconnaissant les cibles avec des affinités de l'ordre du

nono et du micromolaire. La structure des « *repedodies* » a été résolue par cristallographie montrant une architecture typique d'une ossature LRR (Fig. 29) en fer à cheval.

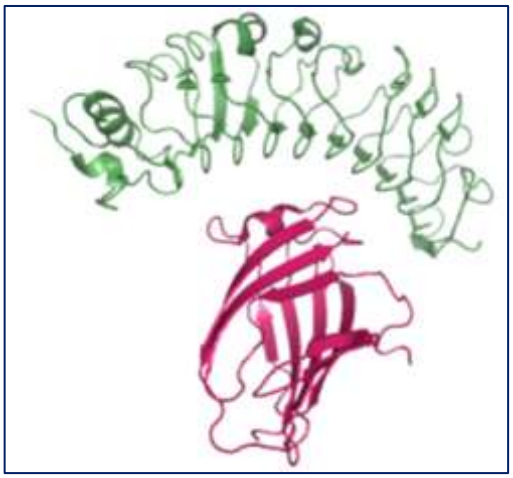


Fig. 29 : Structure du complexe MD2 (en rose) / MD2-*repebody* (en vert) : Le MD2-*repebody* a été conçu à partir de l'ossature de deuxième génération par greffage d'un site d'interaction à partir du site naturel d'un VLR.

L'étape ultime consiste alors en la construction, à partir de cette ossature d'une bibliothèque générique qui permettra de générer des *repebodies* spécifiques à des cibles préalablement choisies. Une bibliothèque a été alors construite en mutant aléatoirement les positions 8, 10 et 11 des motifs $LRRV_1$ et $LRRV_2$. Par *phage display*, l'équipe a sélectionné, à partir de la banque de 10^8 variants, plusieurs *repebodies* qui reconnaissent spécifiquement l'IL-6, avec des affinités entre 48 et 117 nM.

Liste des illustrations

Références bibliographiques

Ahlgren, Sara, Anna Orlova, Helena Wållberg, Monika Hansson, Mattias Sandström, Richard Lewsley, Anders Wennborg, Lars Abrahmsén, Vladimir Tolmachev, and Joachim Feldwisch. 2010. "Targeting of HER2-Expressing Tumors Using 111In-ABY-025, a Second-Generation Affibody Molecule with a Fundamentally Reengineered Scaffold." *Journal of Nuclear Medicine* 51 (7) (July): 1131 –1138. doi:10.2967/jnumed.109.073346.

Ahmad, Zuhaida Asra, Swee Keong Yeap, Abdul Manaf Ali, Wan Yong Ho, Noorjahan Banu Mohamed Alitheen, and Muhajir Hamid. 2012. "scFv Antibody: Principles and Clinical Application." *Clinical and Developmental Immunology* 2012: 1–15. doi:10.1155/2012/980250.

Alcalde, Miguel, Miren Zumárraga, Julio Polaina, Antonio Ballesteros, and Francisco J Plou. 2006. "Combinatorial Saturation Mutagenesis by in Vivo Overlap Extension for the Engineering of Fungal Laccases." *Combinatorial Chemistry & High Throughput Screening* 9 (10) (December): 719–727.

Alder, Matthew N, Igor B Rogozin, Lakshminarayan M Iyer, Galina V Glazko, Max D Cooper, and Zeev Pancer. 2005. "Diversity and Function of Adaptive Immune Receptors in a Jawless Vertebrate." *Science (New York, N.Y.)* 310 (5756) (December 23): 1970–1973. doi:10.1126/science.1119420.

Amstutz, P, P Forrer, C Zahnd, and A Plückthun. 2001. "In Vitro Display Technologies: Novel Developments and Applications." *Current Opinion in Biotechnology* 12 (4) (August): 400–405.

Andrade, Miguel A, Carlo Petosa, Sean I O'Donoghue, Christoph W Müller, and Peer Bork. 2001. "Comparison of ARM and HEAT Protein Repeats." *Journal of Molecular Biology* 309 (1): 1–18. doi:10.1006/jmbi.2001.4624.

Andrade, Miguel A., Carolina Perez-Iratxeta, and Chris P. Ponting. 2001. "Protein Repeats: Structures, Functions, and Evolution." *Journal of Structural Biology* 134 (2-3): 117–131. doi:10.1006/jsbi.2001.4392.

Baker, David. 2006. "Prediction and Design of Macromolecular Structures and Interactions." *Philosophical Transactions of the Royal Society B: Biological Sciences* 361 (1467) (March 29): 459–463. doi:10.1098/rstb.2005.1803.

Barbas, C. F., A. S. Kang, R. A. Lerner, and S. J. Benkovic. 1991. "Assembly of Combinatorial Antibody Libraries on Phage Surfaces: The Gene III Site." *Proceedings of the National Academy of Sciences* 88 (18) (September 15): 7978–7982.

Bartlett, John M S, and David Stirling. 2003. "A Short History of the Polymerase Chain Reaction." *Methods in Molecular Biology (Clifton, N.J.)* 226: 3–6. doi:10.1385/1-59259-384-4:3.

Binder, Uli, Gabriele Matschiner, Ina Theobald, and Arne Skerra. 2010. "High-throughput Sorting of an Anticalin Library via EspP-mediated Functional Display on the Escherichia Coli Cell Surface." *Journal of Molecular Biology* 400 (4): 783–802. doi:10.1016/j.jmb.2010.05.049.

Binz, H.Kaspar, Michael T Stumpp, Patrik Forrer, Patrick Amstutz, and Andreas Plückthun. 2003. "Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins." *Journal of Molecular Biology* 332 (2) (September 12): 489–503. doi:10.1016/S0022-2836(03)00896-9.

Blagodatski, Artem, and Vladimir L Katanaev. 2011. "Technologies of Directed Protein Evolution in Vivo." *Cellular and Molecular Life Sciences: CMLS* 68 (7) (April): 1207–1214. doi:10.1007/s00018-010-0610-5.

Bornscheuer, U T, J Altenbuchner, and H H Meyer. 1999. "Directed Evolution of an Esterase: Screening of Enzyme Libraries Based on pH-indicators and a Growth Assay." *Bioorganic & Medicinal Chemistry* 7 (10) (October): 2169–2173.

Bratkovič, Tomaž. 2009. "Progress in Phage Display: Evolution of the Technique and Its Applications." *Cellular and Molecular Life Sciences: CMLS* (November 15). doi:10.1007/s00018-009-0192-2. http://www.ncbi.nlm.nih.gov/pubmed/19915992.

Brustad, Eric M, and Frances H Arnold. 2011. “Optimizing Non-natural Protein Function with Directed Evolution.” *Current Opinion in Chemical Biology* 15 (2) (April): 201–210. doi:10.1016/j.cbpa.2010.11.020.

Burritt, James B., Mark T. Quinn, Mark A. Jutila, Clifford W. Bond, and Algirdas J. Jesaitis. 1995. “Topological Mapping of Neutrophil Cytochrome b Epitopes with Phage-display Libraries.” *Journal of Biological Chemistry* 270 (28): 16974 –16980. doi:10.1074/jbc.270.28.16974.

Camps, Manel, Jussi Naukkarinen, Ben P. Johnson, and Lawrence A. Loeb. 2003. “Targeted Gene Evolution in Escherichia Coli Using a Highly Error-prone DNA Polymerase I.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (17) (August 19): 9727–9732. doi:10.1073/pnas.1333928100.

Caruthers, M H, S L Beaucage, J W Efcavitch, E F Fisher, M D Matteucci, and Y Stabinsky. 1980. “New Chemical Methods for Synthesizing Polynucleotides.” *Nucleic Acids Symposium Series* (7): 215–223.

Chernomordik, Victor, Moinuddin Hassan, Sang Bong Lee, Rafal Zielinski, Amir Gandjbakhche, and Jacek Capala. 2010. “Quantitative Analysis of Her2 Receptor Expression in Vivo by Near-infrared Optical Imaging.” *Molecular Imaging* 9 (4) (August): 192–200.

Collins, John, Nathalie Horn, Johan WadenbÃ¤ck, and Michael Szardenings. 2001. “Cosmix-plexingÂ®: a Novel Recombinatorial Approach for Evolutionary Selection from Combinatorial Libraries.” *Reviews in Molecular Biotechnology* 74 (4): 317–338. doi:10.1016/S1389-0352(01)00019-8.

Cortajarena, Aitziber L, and Lynne Regan. 2011. “Calorimetric Study of a Series of Designed Repeat Proteins: Modular Structure and Modular Folding.” *Protein Science: A Publication of the Protein Society* 20 (2) (February): 336–340. doi:10.1002/pro.564.

D’Andrea, Luca D, and Lynne Regan. 2003. “TPR Proteins: The Versatile Helix.” *Trends in Biochemical Sciences* 28 (12) (December): 655–662.

Demidov, Vadim V. 2002. “Rolling-circle Amplification in DNA Diagnostics: The Power of Simplicity.” *Expert Review of Molecular Diagnostics* 2 (6) (November): 542–548. doi:10.1586/14737159.2.6.542.

Dimitrov, Ariane, Mélanie Quesnoit, Sandrine Moutel, Isabelle Cantaloube, Christian Poüs, and Franck Perez. 2008. “Detection of GTP-Tubulin Conformation in Vivo Reveals a Role for GTP Remnants in Microtubule Rescues.” *Science* 322 (5906) (November 28): 1353–1356. doi:10.1126/science.1165401.

Dinh, Q, N P Weng, M Kiso, H Ishida, A Hasegawa, and D M Marcus. 1996. “High Affinity Antibodies Against Lex and Sialyl Lex from a Phage Display Library.” *Journal of Immunology (Baltimore, Md.: 1950)* 157 (2) (July 15): 732–738.

Ebersbach, Hilmar, Erik Fiedler, Tanja Scheuermann, Markus Fiedler, Milton T. Stubbs, Carola Reimann, Gabriele Proetzel, Rainer Rudolph, and Ulrike Fiedler. 2007. “Affilin–Novel Binding Molecules Based on Human γ-B-Crystallin, an All β-Sheet Protein.” *Journal of Molecular Biology* 372 (1) (September 7): 172–185. doi:10.1016/j.jmb.2007.06.045.

Ellington, A D, and J W Szostak. 1990. “In Vitro Selection of RNA Molecules That Bind Specific Ligands.” *Nature* 346 (6287) (August 30): 818–822. doi:10.1038/346818a0.

Emond, Stéphane, Philippe Mondon, Sandra Pizzut-Serin, Laurent Douchy, Fabien Crozet, Khalil Bouayadi, Hakim Kharrat, Gabrielle Potocki-Véronèse, Pierre Monsan, and Magali Remaud-Simeon. 2008. “A Novel Random Mutagenesis Approach Using Human Mutagenic DNA Polymerases to Generate Enzyme Variant Libraries.” *Protein Engineering, Design & Selection: PEDS* 21 (4) (April): 267–274. doi:10.1093/protein/gzn004.

Ermakova-Gerdes, S, S Shestakov, and W Vermaas. 1996. “Random Chemical Mutagenesis of a Specific psbDI Region Coding for a Lumenal Loop of the D2 Protein of Photosystem II in Synechocystis Sp. PCC 6803.” *Plant Molecular Biology* 30 (2) (January): 243–254.

Fields, Stanley, and Ok-kyu Song. 1989. “A Novel Genetic System to Detect Protein﷿protein Interactions.” *Nature* 340 (6230) (July): 245–246. doi:10.1038/340245a0.

Fire, A, and S Q Xu. 1995. “Rolling Replication of Short DNA Circles.” *Proceedings of the National Academy of Sciences* 92 (10): 4641 –4645.

Flower, D R. 1996. “The Lipocalin Protein Family: Structure and Function.” *The Biochemical Journal* 318 (Pt 1) (August 15): 1–14.

Fong, H K, J B Hurley, R S Hopkins, R Miake-Lye, M S Johnson, R F Doolittle, and M I Simon. 1986. “Repetitive Segmental Structure of the Transducin Beta Subunit: Homology with the CDC4 Gene and Identification of Related mRNAs.” *Proceedings of the National Academy of Sciences of the United States of America* 83 (7) (April): 2162–2166.

Fujii, Ryota, Motomitsu Kitaoka, and Kiyoshi Hayashi. 2006. “Error-prone Rolling Circle Amplification: The Simplest Random Mutagenesis Protocol.” *Nat. Protocols* 1 (5): 2493–2497. doi:10.1038/nprot.2006.403.

Garcia-Higuera, Irene, Jessica Fenoglio, Ying Li, Carol Lewis, Mikhail P. Panchenko, Orly Reiner, Temple F. Smith, and Eva J. Neer. 1996. “Folding of Proteins with WD-Repeats: Comparison of Six Members of the WD-Repeat Superfamily to the G Protein β Subunit.” *Biochemistry* 35 (44): 13985–13994. doi:10.1021/bi9612879.

Geoffroy, F, R Sodoyer, and L Aujame. 1994. “A New Phage Display System to Construct Multicombinatorial Libraries of Very Large Antibody Repertoires.” *Gene* 151 (1-2) (December 30): 109–113.

Getmanova, Elena V., Yan Chen, Laird Bloom, Jochem Gokemeijer, Steven Shamah, Veena Warikoo, Jack Wang, Vincent Ling, and Lin Sun. 2006. “Antagonists to Human and Mouse Vascular Endothelial Growth Factor Receptor 2 Generated by Directed Protein Evolution In Vitro.” *Chemistry & Biology* 13 (5) (May): 549–556. doi:10.1016/j.chembiol.2005.12.009.

Gill, Pooria, and Amir Ghaemi. 2008. “Nucleic Acid Isothermal Amplification Technologies—A Review.” *Nucleosides, Nucleotides and Nucleic Acids* 27 (3): 224–243. doi:10.1080/15257770701845204.

Groves, Matthew R, and David Barford. 1999. “Topological Characteristics of Helical Repeat Protein.” *Current Opinion in Structural Biology* 9 (3): 383–389. doi:10.1016/S0959-440X(99)80052-9.

Guglielmi, Laurence, Vincent Denis, Nadia Vezzio-Vié, Nicole Bec, Piona Dariavach, Christian Larroque, and Pierre Martineau. 2011. “Selection for Intrabody Solubility in Mammalian Cells Using GFP Fusions.” *Protein Engineering, Design & Selection: PEDS* 24 (12) (December): 873–881. doi:10.1093/protein/gzr049.

Haaparanta, Tapio, and William D. Huse. 1995. “A Combinatorial Method for Constructing Libraries of Long Peptides Displayed by Filamentous Phage.” *Molecular Diversity* 1 (September): 39–52. doi:10.1007/BF01715808.

Hall, B G. 1999. “Experimental Evolution of Ebg Enzyme Provides Clues About the Evolution of Catalysis and to Evolutionary Potential.” *FEMS Microbiology Letters* 174 (1) (May 1): 1–8.

Hanes, Jozef, and Andreas Plückthun. 1997. “In Vitro Selection and Evolution of Functional Proteins by Using Ribosome Display.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (10) (May 13): 4937–4942.

He, M, M Menges, M A Groves, E Corps, H Liu, M Brüggemann, and M J Taussig. 1999. “Selection of a Human Anti-progesterone Antibody Fragment from a Transgenic Mouse Library by ARM Ribosome Display.” *Journal of Immunological Methods* 231 (1-2) (December 10): 105–117.

Hirano, Tatsuya, Noriyuki Kinoshita, Kosuke Morikawa, and Mitsuhiro Yanagida. 1990. “Snap Helix with Knob and Hole: Essential Repeats in S. Pombe Nuclear Protein Nuc2 +.” *Cell* 60 (2) (January 26): 319–328. doi:10.1016/0092-8674(90)90746-2.

Ho, Mitchell, and Ira Pastan. 2009. “Mammalian Cell Display for Antibody Engineering.” *Methods in Molecular Biology (Clifton, N.J.)* 525: 337–352, xiv. doi:10.1007/978-1-59745-554-1_18.

Jackrel, Meredith E., Aitziber L. Cortajarena, Tina Y. Liu, and Lynne Regan. 2009. “Screening Libraries To Identify Proteins with Desired Binding Activities Using a Split-GFP Reassembly Assay.” *ACS Chem. Biol.* 5 (6): 553–562. doi:10.1021/cb900272j.

Jiang, Lei, Zheng Miao, Richard H Kimura, Hongguang Liu, Jennifer R Cochran, Cathy S Culter, Ande Bao, Peiyong Li, and Zhen Cheng. 2011. “Preliminary Evaluation of (177)Lu-labeled Knottin Peptides for Integrin Receptor-targeted Radionuclide Therapy.” *European Journal of Nuclear Medicine and Molecular Imaging* 38 (4) (April): 613–622. doi:10.1007/s00259-010-1684-x.

Kajander, Tommi, Aitziber L Cortajarena, and Lynne Regan. 2006. “Consensus Design as a Tool for Engineering Repeat Proteins.” *Methods in Molecular Biology (Clifton, N.J.)* 340: 151–170. doi:10.1385/1-59745-116-9:151.

Kanayama, Naoki, Kagefumi Todo, Satoko Takahashi, Masaki Magari, and Hitoshi Ohmori. 2006. “Genetic Manipulation of an Exogenous Non-immunoglobulin Protein by Gene Conversion Machinery in a Chicken B Cell Line.” *Nucleic Acids Research* 34 (2): e10–e10. doi:10.1093/nar/gnj013.

Kasahara, N, A M Dozy, and Y W Kan. 1994. “Tissue-specific Targeting of Retroviral Vectors Through Ligand-receptor Interactions.” *Science (New York, N.Y.)* 266 (5189) (November 25): 1373–1376.

Kashiwagi, Kenji, Yasuhiro Isogai, Kei-ichi Nishiguchi, and Kiyotaka Shiba. 2006. “Frame Shuffling: a Novel Method for in Vitro Protein Evolution.” *Protein Engineering, Design & Selection: PEDS* 19 (3) (March): 135–140. doi:10.1093/protein/gzj008.

Keefe, A D, and J W Szostak. 2001. “Functional Proteins from a Random-sequence Library.” *Nature* 410 (6829) (April 5): 715–718. doi:10.1038/35070613.

Kieke, M C, B K Cho, E T Boder, D M Kranz, and K D Wittrup. 1997. “Isolation of anti-T Cell Receptor scFv Mutants by Yeast Surface Display.” *Protein Engineering* 10 (11) (November): 1303–1310.

Kimura, Richard H., Aron M. Levin, Frank V. Cochran, and Jennifer R. Cochran. 2009. “Engineered Cystine Knot Peptides That Bind Avβ3, Avβ5, and A5β1 Integrins with Low-nanomolar Affinity.” *Proteins: Structure, Function, and Bioinformatics* 77 (2) (November 1): 359–369. doi:10.1002/prot.22441.

Knappik, A, L Ge, A Honegger, P Pack, M Fischer, G Wellnhofer, A Hoess, J Wölle, A Plückthun, and B Virnekäs. 2000. “Fully Synthetic Human Combinatorial Antibody Libraries (HuCAL) Based on Modular Consensus Frameworks and CDRs Randomized with Trinucleotides.” *Journal of Molecular Biology* 296 (1) (February 11): 57–86. doi:10.1006/jmbi.1999.3444.

Kobe, B, and J Deisenhofer. 1995. “Proteins with Leucine-rich Repeats.” *Current Opinion in Structural Biology* 5 (3) (June): 409–416.

Kobe, B, and A V Kajava. 2001. “The Leucine-rich Repeat as a Protein Recognition Motif.” *Current Opinion in Structural Biology* 11 (6) (December): 725–732.

Koide, Akiko, John Wojcik, Ryan N Gilbreth, Annett Reichel, Jacob Piehler, and Shohei Koide. 2009. “Accelerating Phage-display Library Selection by Reversible and Site-specific Biotinylation.” *Protein Engineering, Design & Selection: PEDS* 22 (11) (November): 685–690. doi:10.1093/protein/gzp053.

Kondo, A, and M Ueda. 2004. “Yeast Cell-surface Display--applications of Molecular Display.” *Applied Microbiology and Biotechnology* 64 (1) (March): 28–40. doi:10.1007/s00253-003-1492-3.

Kramer, R Arjen, Freek Cox, Marieke van der Horst, Sonja van der Oudenrijn, Pieter C M Res, Judith Bia, Ton Logtenberg, and John de Kruif. 2003. “A Novel Helper Phage That Improves Phage Display Selection Efficiency by Preventing the Amplification of Phages Without Recombinant Protein.” *Nucleic Acids Research* 31 (11) (June 1): e59.

Kronqvist, Nina, Magdalena Malm, Lovisa Göstring, Elin Gunneriusson, Martin Nilsson, Ingmarie Höidén Guthenberg, Lars Gedda, Fredrik Y Frejd, Stefan Ståhl, and John Löfblom. 2011. “Combining Phage and Staphylococcal Surface Display for Generation of ErbB3-specific Affibody Molecules.” *Protein Engineering, Design & Selection: PEDS* 24 (4) (April): 385–396. doi:10.1093/protein/gzq118.

Krüger DH, Bickle TA, and Krüger DH, Bickle TA. 1983. “Bacteriophage Survival: Multiple Mechanisms for Avoiding the Deoxyribonucleic Acid Restriction Systems of Their Hosts.” *Microbiol. Rev.* 47 (3) (September): 345–60.

Labrou, Nikolaos E. 2010a. “Random Mutagenesis Methods for in Vitro Directed Enzyme Evolution.” *Current Protein & Peptide Science* 11 (1) (February 1): 91–100.

———. 2010b. “Random Mutagenesis Methods for in Vitro Directed Enzyme Evolution.” *Current Protein & Peptide Science* 11 (1) (February 1): 91–100.

Lee, Chang-Han, Kyung-Jin Park, Eun-Sil Sung, Aeyung Kim, Ji-Da Choi, Jeong-Sun Kim, Soo-Hyun Kim, Myung-Hee Kwon, and Yong-Sung Kim. 2010. "Engineering of a Human Kringle Domain into Agonistic and Antagonistic Binding Proteins Functioning in Vitro and in Vivo." *Proceedings of the National Academy of Sciences of the United States of America* 107 (21) (May 25): 9567–9571. doi:10.1073/pnas.1001541107.

Lee, Jihun, Sachiko I. Blaber, Vikash K. Dubey, and Michael Blaber. 2011. "A Polypeptide 'Building Block' for the β-Trefoil Fold Identified by 'Top-Down Symmetric Deconstruction'." *Journal of Molecular Biology* 407 (5) (April 15): 744–763. doi:10.1016/j.jmb.2011.02.002.

Lee, Sang-Chul, Keunwan Park, Jieun Han, Joong-jae Lee, Hyun Jung Kim, Seungpyo Hong, Woosung Heu, et al. 2012. "Design of a Binding Scaffold Based on Variable Lymphocyte Receptors of Jawless Vertebrates by Module Engineering." *Proceedings of the National Academy of Sciences of the United States of America* 109 (9) (February 28): 3299–3304. doi:10.1073/pnas.1113193109.

Lipovsek, D. 2011. "Adnectins: Engineered Target-binding Protein Therapeutics." *Protein Engineering, Design & Selection: PEDS* 24 (1-2) (January): 3–9. doi:10.1093/protein/gzq097.

Löfblom, J, J Feldwisch, V Tolmachev, J Carlsson, S Ståhl, and F Y Frejd. 2010. "Affibody Molecules: Engineered Proteins for Therapeutic, Diagnostic and Biotechnological Applications." *FEBS Letters* 584 (12) (June 18): 2670–2680. doi:10.1016/j.febslet.2010.04.014.

Löfblom, John, Fredrik Y Frejd, and Stefan Ståhl. 2011. "Non-immunoglobulin Based Protein Scaffolds." *Current Opinion in Biotechnology* (July 2). doi:10.1016/j.copbio.2011.06.002. http://www.ncbi.nlm.nih.gov/pubmed/21726995.

Maeda, Toshinari, Viviana Sanchez-Torres, and Thomas K. Wood. 2008. "Protein Engineering of Hydrogenase 3 to Enhance Hydrogen Production." *Applied Microbiology and Biotechnology* 79 (1) (March): 77–86. doi:10.1007/s00253-008-1416-3.

Magliery, Thomas J, and Lynne Regan. "Sequence Variation in Ligand Binding Sites in Proteins." *BMC Bioinformatics* 6: 240–240. doi:10.1186/1471-2105-6-240.

Magliery, Thomas J., and Lynne Regan. 2004. "Beyond Consensus: Statistical Free Energies Reveal Hidden Interactions in the Design of a TPR Motif." *Journal of Molecular Biology* 343 (3) (October 22): 731–745. doi:10.1016/j.jmb.2004.08.026.

Main, Ewan R.G., Yong Xiong, Melanie J. Cocco, Luca D'Andrea, and Lynne Regan. 2003. "Design of Stable α-Helical Arrays from an Idealized TPR Motif." *Structure* 11 (5): 497–508. doi:10.1016/S0969-2126(03)00076-5.

de Marco, Ario. 2009. "Strategies for Successful Recombinant Expression of Disulfide Bond-dependent Proteins in Escherichia Coli." *Microbial Cell Factories* 8: 26. doi:10.1186/1475-2859-8-26.

Matsumura, Ichiro, and Lori A Rowe. 2005. "Whole Plasmid Mutagenic PCR for Directed Protein Evolution." *Biomolecular Engineering* 22 (1-3) (June): 73–79. doi:10.1016/j.bioeng.2004.10.004.

Matsumura, Nobutaka, Toru Tsuji, Takeshi Sumida, Masahito Kokubo, Michiko Onimaru, Nobuhide Doi, Hideaki Takashima, Etsuko Miyamoto-Sato, and Hiroshi Yanagawa. 2010. "mRNA Display Selection of a High-affinity, Bcl-X(L)-specific Binding Peptide." *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 24 (7) (July): 2201–2210. doi:10.1096/fj.09-143008.

Matsuura, Tomoaki, and Tetsuya Yomo. 2006. "In Vitro Evolution of Proteins." *Journal of Bioscience and Bioengineering* 101 (6): 449–456. doi:10.1263/jbb.101.449.

Mattheakis, L C, R R Bhatt, and W J Dower. 1994. "An in Vitro Polysome Display System for Identifying Ligands from Very Large Peptide Libraries." *Proceedings of the National Academy of Sciences of the United States of America* 91 (19) (September 13): 9022–9026.

McConnell, S J, A J Uveges, D M Fowlkes, and D G Spinella. 1996. "Construction and Screening of M13 Phage Libraries Displaying Long Random Peptides." *Molecular Diversity* 1 (3) (May): 165–176.

McLafferty, M A, R B Kent, R C Ladner, and W Markland. 1993. "M13 Bacteriophage Displaying Disulfide-constrained Microproteins." *Gene* 128 (1) (June 15): 29–36.

Miao, Zheng, Gang Ren, Hongguang Liu, Richard H Kimura, Lei Jiang, Jennifer R Cochran, Sanjiv Sam Gambhir, and Zhen Cheng. 2009. "An Engineered Knottin Peptide Labeled with 18F for PET Imaging of Integrin Expression." *Bioconjugate Chemistry* 20 (12) (December): 2342–2347. doi:10.1021/bc900361g.

Michnick, Stephen W. 2003. "Protein Fragment Complementation Strategies for Biochemical Network Mapping." *Current Opinion in Biotechnology* 14 (6) (December): 610–617.

Michnick, Stephen W. 2001. "Exploring Protein Interactions by Interaction-induced Folding of Proteins from Complementary Peptide Fragments." *Current Opinion in Structural Biology* 11 (4): 472–477. doi:16/S0959-440X(00)00235-9.

Minenkova, Olga, Andrea Pucci, Emiliano Pavoni, Amedeo De Tomassi, Paola Fortugno, Nicola Gargano, Maurizio Cianfriglia, et al. 2003. "Identification of Tumor-associated Antigens by Screening Phage-displayed Human cDNA Libraries with Sera from Tumor Patients." *International Journal of Cancer* 106 (4): 534–544. doi:10.1002/ijc.11269.

Mirecka, Ewa A, Thomas Hey, Ulrike Fiedler, Rainer Rudolph, and Mechthild Hatzfeld. 2009. "Affilin Molecules Selected Against the Human Papillomavirus E7 Protein Inhibit the Proliferation of Target Cells." *Journal of Molecular Biology* 390 (4) (July 24): 710–721. doi:10.1016/j.jmb.2009.05.027.

Mohan, Utpal, and Uttam Chand Banerjee. 2008. "Molecular Evolution of a Defined DNA Sequence with Accumulation of Mutations in a Single Round by a Dual Approach to Random Chemical Mutagenesis (DuARCheM)." *ChemBioChem* 9 (14) (September 22): 2238–2243. doi:10.1002/cbic.200800259.

Mosavi, Leila K., Daniel L. Minor, and Zheng-yu Peng. 2002. "Consensus-derived Structural Determinants of the Ankyrin Repeat Motif." *Proceedings of the National Academy of Sciences* 99 (25): 16029 –16034. doi:10.1073/pnas.252537899.

Moutel, Sandrine, and Franck Perez. 2009. "Utilisation Des Intrabodies : De L'étude Des Protéines Intracellulaires à L'immunisation Thérapeutique." *Médecine/sciences* 25 (12) (December 15): 1173–1176. doi:10.1051/medsci/200925121173.

Munch, Robert C, Michael D Muhlebach, Thomas Schaser, Sabrina Kneissl, Christian Jost, Andreas Pluckthun, Klaus Cichutek, and Christian J Buchholz. 2011. "DARPins: An Efficient Targeting Domain for Lentiviral Vectors." *Mol Ther* 19 (4): 686–693.

Nakashima, Nobutaka, and Tomohiro Tamura. 2009. "Conditional Gene Silencing of Multiple Genes with Antisense RNAs and Generation of a Mutator Strain of Escherichia Coli." *Nucleic Acids Research* 37 (15) (August): e103. doi:10.1093/nar/gkp498.

Nord, K, E Gunneriusson, J Ringdahl, S Ståhl, M Uhlén, and P A Nygren. 1997. "Binding Proteins Selected from Combinatorial Libraries of an Alpha-helical Bacterial Receptor Domain." *Nature Biotechnology* 15 (8) (August): 772–777. doi:10.1038/nbt0897-772.

Olsen, M J, D Stephens, D Griffiths, P Daugherty, G Georgiou, and B L Iverson. 2000. "Function-based Isolation of Novel Enzymes from a Large Library." *Nature Biotechnology* 18 (10) (October): 1071–1074. doi:10.1038/80267.

Orlova, Anna, Helena Wållberg, Sharon Stone-Elander, and Vladimir Tolmachev. 2009. "On the Selection of a Tracer for PET Imaging of HER2-Expressing Tumors: Direct Comparison of a 124I-Labeled Affibody Molecule and Trastuzumab in a Murine Xenograft Model." *Journal of Nuclear Medicine* 50 (3) (March): 417 –425. doi:10.2967/jnumed.108.057919.

Pande, Jyoti, Magdalena M Szewczyk, and Ashok K Grover. 2010. "Phage Display: Concept, Innovations, Applications and Future." *Biotechnology Advances* 28 (6) (December): 849–858. doi:10.1016/j.biotechadv.2010.07.004.

Pande, Jyoti, Magdalena M Szewczyk, Iwona Kuszczak, Shawn Grover, E Escher, and Ashok K Grover. 2008. "Functional Effects of Caloxin 1c2, a Novel Engineered Selective Inhibitor of Plasma Membrane Ca(2+)-pump Isoform 4, on Coronary Artery." *Journal of Cellular and Molecular Medicine* 12 (3) (June): 1049–1060. doi:10.1111/j.1582-4934.2008.00140.x.

Park, Jong Pil, Donald M Cropek, and Scott Banta. 2010. “High Affinity Peptides for the Recognition of the Heart Disease Biomarker Troponin I Identified Using Phage Display.” *Biotechnology and Bioengineering* 105 (4) (March 1): 678–686. doi:10.1002/bit.22597.

Peifer, Mark, Sven Berg, and Albert B. Reynolds. 1994. “A Repeating Amino Acid Motif Shared by Proteins with Diverse Cellular Roles.” *Cell* 76 (5) (March 11): 789–791. doi:10.1016/0092-8674(94)90353-0.

Pirakitikulr, Nathan, Nili Ostrov, Pamela Peralta-Yahya, and Virginia W Cornish. 2010. “PCRless Library Mutagenesis via Oligonucleotide Recombination in Yeast.” *Protein Science* 19 (12) (December 1): 2336–2346. doi:10.1002/pro.513.

Ponting, Chris P, and Robert B Russell. 2000. “Identification of Distant Homologues of Fibroblast Growth Factors Suggests a Common Ancestor for All B-trefoil Proteins.” *Journal of Molecular Biology* 302 (5) (October 6): 1041–1047. doi:10.1006/jmbi.2000.4087.

Ren, Zhao-jun, and Lindsay W Black. 1998. “Phage T4 SOC and HOC Display of Biologically Active, Full-length Proteins on the Viral Capsid.” *Gene* 215 (2) (July 30): 439–444. doi:10.1016/S0378-1119(98)00298-4.

Roberts, R W, and J W Szostak. 1997. “RNA-peptide Fusions for the in Vitro Selection of Peptides and Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (23) (November 11): 12297–12302.

Rondot, S, J Koch, F Breitling, and S Dübel. 2001. “A Helper Phage to Improve Single-chain Antibody Presentation in Phage Display.” *Nature Biotechnology* 19 (1) (January): 75–78. doi:10.1038/83567.

Sale, Julian E, and Michael S Neuberger. 1998. “TdT-Accessible Breaks Are Scattered over the Immunoglobulin V Domain in a Constitutively Hypermutating B Cell Line.” *Immunity* 9 (6) (December): 859–869. doi:10.1016/S1074-7613(00)80651-2.

von Schantz, Laura, Fredrika Gullfot, Sebastian Scheer, Lada Filonova, Lavinia Cicortas Gunnarsson, James E Flint, Geoffrey Daniel, Eva Nordberg-Karlsson, Harry Brumer, and Mats Ohlin. “Affinity Maturation Generates Greatly Improved Xyloglucan-specific Carbohydrate Binding Modules” 9: 92–92. doi:10.1186/1472-6750-9-92.

Schlehuber, Steffen, and Arne Skerra. 2005. “Lipocalins in Drug Discovery: From Natural Ligand-binding Proteins to ‘Anticalins’.” *Drug Discovery Today* 10 (1) (January 1): 23–33. doi:10.1016/S1359-6446(04)03294-5.

Sedgwick, Steven G, and Stephen J Smerdon. 1999. “The Ankyrin Repeat: a Diversity of Interactions on a Common Structural Framework.” *Trends in Biochemical Sciences* 24 (8): 311–316. doi:10.1016/S0968-0004(99)01426-7.

Seo, Hidetaka, Shu-ichi Hashimoto, Kyoko Tsuchiya, Waka Lin, Takehiko Shibata, and Kunihiro Ohta. 2006. “An Ex Vivo Method for Rapid Generation of Monoclonal Antibodies (ADLib System).” *Nat. Protocols* 1 (3) (November): 1502–1506. doi:10.1038/nprot.2006.248.

Sheedy, Claudia, C Roger MacKenzie, and J Christopher Hall. 2007. “Isolation and Affinity Maturation of Hapten-specific Antibodies.” *Biotechnology Advances* 25 (4) (August): 333–352. doi:10.1016/j.biotechadv.2007.02.003.

Shibasaki, Seiji, and Mitsuyoshi Ueda. 2010. “Development of Yeast Molecular Display Systems Focused on Therapeutic Proteins, Enzymes, and Foods: Functional Analysis of Proteins and Its Application to Bioconversion.” *Recent Patents on Biotechnology* 4 (3) (November): 198–213.

Sidhu, S S. 2001. “Engineering M13 for Phage Display.” *Biomolecular Engineering* 18 (2) (September): 57–63.

Smith, Colin A., and Tanja Kortemme. “Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design” 6 (7). doi:10.1371/journal.pone.0020451.

Smith, G P. 1985. “Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface.” *Science (New York, N.Y.)* 228 (4705) (June 14): 1315–1317.

Smith, George P., and Valery A. Petrenko. 1997. “Phage Display.” *Chemical Reviews* 97 (2) (April 1): 391–410.

Smith, Temple F, Chrysanthe Gaitatzes, Kumkum Saxena, and Eva J Neer. 1999. "The WD Repeat: a Common Architecture for Diverse Functions." *Trends in Biochemical Sciences* 24 (5): 181–185. doi:10.1016/S0968-0004(99)01384-5.

Sommerhoff, Christian P., Olga Avrutina, Hans-Ulrich Schmoldt, Dusica Gabrijelcic-Geiger, Ulf Diederichsen, and Harald Kolmar. 2010. "Engineered Cystine Knot Miniproteins as Potent Inhibitors of Human Mast Cell Tryptase β." *Journal of Molecular Biology* 395 (1) (January 8): 167–175. doi:10.1016/j.jmb.2009.10.028.

Stoltenburg, Regina, Christine Reinemann, and Beate Strehlitz. 2007. "SELEX--a (r)evolutionary Method to Generate High-affinity Nucleic Acid Ligands." *Biomolecular Engineering* 24 (4) (October): 381–403. doi:10.1016/j.bioeng.2007.06.001.

Sumida, Takeshi, Nobuhide Doi, and Hiroshi Yanagawa. 2009. "Bicistronic DNA Display for in Vitro Selection of Fab Fragments." *Nucleic Acids Research* 37 (22) (December): e147. doi:10.1093/nar/gkp776.

Takahashi, N, Y Takahashi, and F W Putnam. 1985. "Periodicity of Leucine and Tandem Repetition of a 24-amino Acid Segment in the Primary Structure of Leucine-rich Alpha 2-glycoprotein of Human Serum." *Proceedings of the National Academy of Sciences of the United States of America* 82 (7) (April): 1906–1910.

Tawfik, D S, and A D Griffiths. 1998. "Man-made Cell-like Compartments for Molecular Evolution." *Nature Biotechnology* 16 (7) (July): 652–656. doi:10.1038/nbt0798-652.

Theurillat, Jean-Philippe, Birgit Dreier, Gabriela Nagy-Davidescu, Burkhardt Seifert, Silvia Behnke, Ursina Zurrer-Hardi, Fabienne Ingold, Andreas Pluckthun, and Holger Moch. 2010. "Designed Ankyrin Repeat Proteins: a Novel Tool for Testing Epidermal Growth Factor Receptor 2 Expression in Breast Cancer." *Mod Pathol* 23 (9): 1289–1297.

Thie, Holger, Bernd Voedisch, Stefan Dübel, Michael Hust, and Thomas Schirrmann. 2009. "Affinity Maturation by Phage Display." In *Therapeutic Antibodies*, ed. Antony S. Dimitrov, 525:309–322. Totowa, NJ: Humana Press. http://www.springerlink.com.gate1.inist.fr/content/l232242669467g17/#section=54892&page=1&locus=98.

Tolcher, Anthony W, Christopher J Sweeney, Kyri Papadopoulos, Amita Patnaik, Elena G Chiorean, Alain C Mita, Kamalesh Sankhala, et al. 2011. "Phase I and Pharmacokinetic Study of CT-322 (BMS-844203), a Targeted Adnectin Inhibitor of VEGFR-2 Based on a Domain of Human Fibronectin." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 17 (2) (January 15): 363–371. doi:10.1158/1078-0432.CCR-10-1411.

Urvoas, Agathe, Asma Guellouz, Marie Valerio-Lepiniec, Marc Graille, Dominique Durand, Danielle C. Desravines, Herman van Tilbeurgh, Michel Desmadril, and Philippe Minard. 2010a. "Design, Production and Molecular Structure of a New Family of Artificial Alpha-helicoidal Repeat Proteins (αRep) Based on Thermostable HEAT-like Repeats." *Journal of Molecular Biology* 404 (2) (November 26): 307–327. doi:10.1016/j.jmb.2010.09.048.

———. 2010b. "Design, Production and Molecular Structure of a New Family of Artificial Alpha-helicoidal Repeat Proteins ([alpha]Rep) Based on Thermostable HEAT-like Repeats." *Journal of Molecular Biology* 404 (2) (November 26): 307–327. doi:16/j.jmb.2010.09.048.

Varghese, J N, J L McKimm-Breschkin, J B Caldwell, A A Kortt, and P M Colman. 1992. "The Structure of the Complex Between Influenza Virus Neuraminidase and Sialic Acid, the Viral Receptor." *Proteins* 14 (3) (November): 327–332. doi:10.1002/prot.340140302.

Velikovsky, C Alejandro, Lu Deng, Satoshi Tasumi, Lakshminarayan M Iyer, Melissa C Kerzic, L Aravind, Zeev Pancer, and Roy A Mariuzza. 2009. "Structure of a Lamprey Variable Lymphocyte Receptor in Complex with a Protein Antigen." *Nature Structural & Molecular Biology* 16 (7) (July): 725–730. doi:10.1038/nsmb.1619.

Verschueren, Erik, Peter Vanhee, Almer M van der Sloot, Luis Serrano, Frederic Rousseau, and Joost Schymkowitz. 2011. "Protein Design with Fragment Databases." *Current Opinion in Structural Biology* 21 (4): 452–459. doi:10.1016/j.sbi.2011.05.002.

Vieira, Jeffrey, and Joachim Messing. 1987. "[1] Production of Single-stranded Plasmid DNA." In *Recombinant DNA Part D*, Volume 153:3–11. Academic Press. http://www.sciencedirect.com/science/article/pii/0076687987530440.

Wang, Harris H, Farren J Isaacs, Peter A Carr, Zachary Z Sun, George Xu, Craig R Forest, and George M Church. 2009. "Programming Cells by Multiplex Genome Engineering and Accelerated Evolution." *Nature* 460 (7257) (August 13): 894–898. doi:10.1038/nature08187.

Waterhouse, P, A D Griffiths, K S Johnson, and G Winter. 1993. "Combinatorial Infection and in Vivo Recombination: a Strategy for Making Large Phage Antibody Repertoires." *Nucleic Acids Research* 21 (9) (May 11): 2265–2266.

Winkler, Johannes, Patricia Martin-Killias, Andreas Plückthun, and Uwe Zangemeister-Wittke. 2009. "EpCAM-targeted Delivery of Nanocomplexed siRNA to Tumor Cells with Designed Ankyrin Repeat Proteins." *Molecular Cancer Therapeutics* 8 (9): 2674 –2683. doi:10.1158/1535-7163.MCT-09-0402.

Yamaguchi, Junichi, Mohammed Naimuddin, Manish Biyani, Toru Sasaki, Masayuki Machida, Tai Kubo, Takashi Funatsu, Yuzuru Husimi, and Naoto Nemoto. 2009. "cDNA Display: a Novel Screening Method for Functional Disulfide-rich Peptides by Solid-phase Synthesis and Stabilization of mRNA–protein Fusions" 37 (16) (September): e108–e108. doi:10.1093/nar/gkp514.

Chapitre I :

Conception et création d'une banque de protéines artificielles : les αRep.

Table des matières

Introduction

Nous avons passé en revue dans le chapitre précédent, des travaux pertinents dans le domaine de la création de banques de protéines artificielles à partir d'ossatures protéiques naturelles. En effet, les travaux innovants et les résultats avancés obtenus par l'équipe de Plückthun avec les *DARPins* et celle de Regan avec les TPR et les *Repeatbodies* de Lee démontraient clairement que ce type d'ossature ouvrait des perspectives larges.

L'objectif que s'était fixé notre laboratoire depuis plusieurs années, était précisément de développer des méthodes permettant de créer efficacement des sites de fixation choisis dans des protéines. Il semblait donc important de contribuer à explorer ces ossatures prometteuses. Le laboratoire avait exploré durant plusieurs années, sans parvenir à des résultats très encourageants, une ossature de type LRR (*Leucine rich repeats*) ce qui avait permis d'entrevoir les avantages mais surtout les difficultés spécifiques de ces protéines particulières.

Nous nous sommes alors intéressés à une autre famille d'ossatures protéiques à motifs répétés, non encore exploitée dans cette perspective, qui est la famille des *Heat repeats*. Le travail a démarré à partir d'une recherche systématique, dans la banque de données structurales d'ossatures protéiques pouvant servir de base lors de la conception des protéines artificielles. L'identification de la protéine ayant servi de guide, Mth187, et le travail de définition d'un consensus avait été réalisés préalablement à mon arrivée au laboratoire. De même, les technologies de construction de banque de *repeats* avaient été explorées et une banque de première génération de ce qui allait devenir les *αRep* venait d'être construite. Cette banque comprend $3*10^8$ clones indépendants dont les séquences codent des protéines ayant toutes la même architecture générale reposant sur la répétition d'un motif constitué de deux hélices α. Les protéines de cette banque comportent une séquence différente dans les positions hypervariables de chaque motif répété et un nombre de motifs insérés différent d'une protéine à une autre. Cette banque avait été surtout construite pour vérifier si la conception des protéines était correcte, et si la tolérance supposée à la variabilité en taille et en séquence était bien réelle.

Mon premier objectif, lorsque j'ai intégré l'équipe Modélisation et ingénierie des protéines, était de caractériser cette banque et les protéines qu'elle recelait : les étapes étaient successivement : l'analyse statistique des séquences de clones pris au hasard de la banque, du pourcentage de clones codants de cette banque et en particulier exprimés solubles, la vérification de la structure secondaire et de la stabilité thermique des variants de la banque par

rapport au modèle conçu au départ. Ma contribution à cette étape du travail s'est faite avec une collaboration quotidienne avec Agathe Urvoas et Marielle Valerio. Ce travail, qui fait objet de ce premier chapitre du manuscrit de thèse, nous a permis de publier dans Journal of Molecular Biology (Guellouz et *al.*, 2010).

Design, Production and Molecular Structure of a New Family of Artificial Alpha-helicoïdal Repeat Proteins (αRep) Based on Thermostable HEAT-like Repeat.

Agathe Urvoas[1, 2, †], Asma Guellouz[1, 2, †], Marie Valerio-Lepiniec[1, 2], Marc Graille[1, 2], Dominique Durand[1, 2], Danielle C. Desravines[1, 2, 2], Herman van Tilbeurgh[1, 2], Michel Desmadril[1, 2], Philippe Minard[1, 2].

[1] équipe "modélisation et d'ingénierie des protéines", Institut de Biochimie et Biophysique Moléculaire et Cellulaire, UMR 8619 CNRS, Université Paris Sud 11, 91405 Orsay cedex (France).

[2] équipe "génomique structurale de la levure", Institut de Biochimie et Biophysique Moléculaire et Cellulaire, UMR 8619 CNRS, Université Paris Sud 11, 91405 Orsay cedex (France).

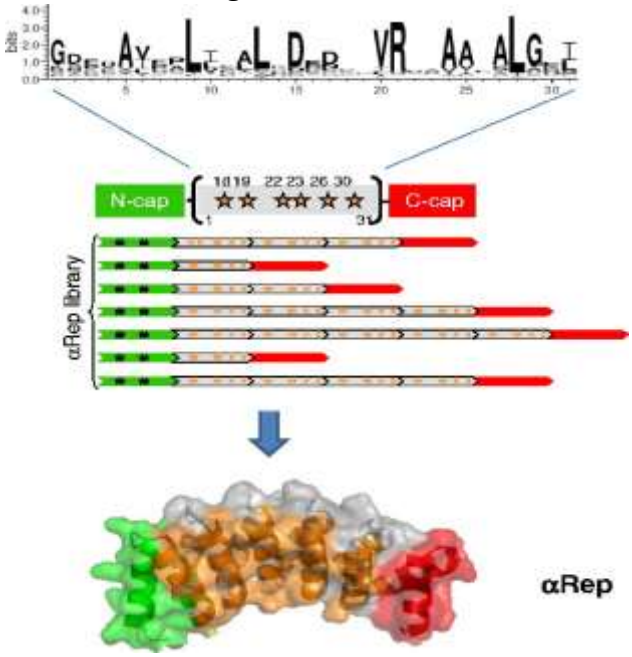
[†] *These authors contributed equally to this work*

Received 16 February 2010; revised 15 September 2010; Accepted 21 September 2010. Edited by F. Schmid. Available online 29 September 2010.

http://www.sciencedirect.com.gate1.inist.fr/science/article/pii/S0022283610010478

Abstract

Repeat proteins have a modular organization and a regular architecture that make them attractive models for design and directed evolution experiments. HEAT repeat proteins, although very common, have not been used as a scaffold for artificial proteins, probably because they are made of long and irregular repeats. Here, we present and validate a consensus sequence for artificial HEAT repeat proteins. The sequence was defined from the structure-based sequence analysis of a thermostable HEAT-like repeat protein. Appropriate sequences were identified for the N- and C-caps. A library of genes coding for artificial proteins based on this sequence design, named αRep, was assembled using new and versatile methodology based on circular amplification. Proteins picked randomly from this library are expressed as soluble proteins. The biophysical properties of proteins with different numbers of repeats and different combinations of side chains in hypervariable positions were characterized. Circular dichroism and differential scanning calorimetry experiments showed that all these proteins are folded cooperatively and are very stable (T_m > 70 °C). Stability of these proteins increases with the number of repeats. Detailed gel filtration and small-angle X-ray scattering studies showed that the purified proteins form either monomers or dimers. The X-ray structure of a stable dimeric variant structure was solved. The protein is folded with a highly regular topology and the repeat structure is organized, as expected, as pairs of alpha helices. In this protein variant, the dimerization interface results directly from the variable surface enriched in aromatic residues located in the randomized positions of the repeats. The dimer was crystallized both in an apo and in a PEG-bound form, revealing a very well defined binding crevice and some structure flexibility at the interface. This fortuitous binding site could later prove to be a useful binding site for other low molecular mass partners.



Research Highlights

► A new family of artificial proteins, named αRep, is described. ► These proteins are made by a repeated two helices motif, designed from thermostable HEAT–like repeats. ► Variable positions on each motif generate a hypervariable macrosurface on the protein. ► Randomly selected αRep proteins are well expressed, folded and very stable. ► The experimental 3D structure of one αRep protein is described.

Keywords: *HEAT* repeat protein; protein design; combinatorial library; protein scaffold

Abbreviations used: DSC, differential scanning calorimetry; SAXS, small-angle X-ray scattering; SEC, size-exclusion chromatography; TPR, tetratricopeptide repeat; RCA, rolling circle amplification; MPD, 2-methyl-2,4-pentanediol

Article Outline

- Synthesis procedures
- Biophysical properties
- Oligomeric structure
- Applications

- Materials and Methods
  - Sequence analysis
  - Expression vectors
    - ✓ Synthesis of the microgenes
  - Library construction
  - Monitoring for soluble expression
  - Protein expression and purification
  - Circular dichroism measurements
  - Thermal denaturation measurements
  - Differential scanning calorimetry (DSC)
  - Analytical size-exclusion chromatography
  - Crystallization and resolution of the structure
  - Small-angle X-ray scattering
    - ✓ SAXS experiments
    - ✓ Data analysis
  - Protein Data Bank accession numbers

Acknowledgements

References

Introduction

Antibodies are, by far, the most commonly used specific binding reagents and are still often seen as the only available option when a tight binding protein of defined specificity is needed. However, the introduction of binding sites of desired specificity has become feasible in other types of proteins not related to antibodies [1], [2] and [3]. These so-called alternative scaffold proteins were developed to overcome the limits of antibodies and derivatives, and notably the poor expression and biophysical properties of most natural antibodies. Proteins with very different topologies have been described as potentially useful scaffolds for engineering. Bacterial helical bundle domains (affibodies),[4] fibronectin modules (monobodies),[5] lipocalin (anticalins)[6] and crystalline domains (affilins),[7] all generated specific binders against predefined protein targets. Our group has developed drug-carrying proteins based on neocarzinostatin. [8] , [9] and [10] Comprehensive coverage of alternative scaffolds has been published. [1] , [2] and [3] These examples demonstrate that the versatile molecular recognition capability of antibodies is not the privilege of the immunoglobulin fold but is, in fact, shared by many different protein architectures.

A very promising route toward versatile alternative scaffolds originated from the idea of creating new artificial proteins based on the principles observed in natural repeat proteins, [11] and [12] such as ankyrin or leucine-rich repeats proteins (LRRs). Repeat proteins are made by the repetition of a simple structural motif.[13] Each motif is 20 – 40 amino acids long, depending on the protein family, and encodes a simple arrangement of a few secondary structure elements. Although a single motif is not able to fold by itself, the consecutive motifs stack on each other to give rise to a protein with an elongated shape and a solenoid-like topology. Multi-alignment of all motifs within a repeat family shows that some positions of the motif are highly conserved and others are highly variable. The conserved positions correspond to residues involved in maintenance of the structure of each motif and/or interaction with neighboring motifs. Variable residues are less likely to contribute to fold stability and cluster on the surface of the repeat proteins, creating a hypervariable macro-surface from which interactions with other protein partners can emerge. These proteins provide versatile scaffolds for molecular recognition and were recruited by natural evolution for a wide range of functions. A remarkable illustration of these versatile molecular recognition abilities was the discovery of the adaptative immune system of Jawless vertebrates, which is not based on the immunoglobulin fold but rather on the LRR family.[14] It

seems plausible that the molecular organization of these repeated proteins easily generates new evolutive trajectories by the duplication/recombination/diversification of simple pre-existing motifs and this might explain their evolutionary success. These natural evolutionary processes can now be reconstituted *in vitro* by creating artificial repeat protein libraries based on consensus repeat sequences. [3] and [15]

Several well-known protein repeat families were studied in order to define an operational set of rules to create artificial repeat proteins. Ankyrin repeat proteins were the first successful family that generated consensus designed new artificial proteins. [15] and [16] Designed artificial ankyrins express well, are extremely stable and adopt a fold identical with that of natural ankyrin proteins. Very diverse repertoires of designed ankyrin repeat proteins (Darpins) were selected by phage or ribosome display, resulting in tight and highly specific binders against a range of different protein targets,[17] including integral membrane proteins. [18] and [19] A sub-family of the LRRs based on the ribonuclease inhibitor has been used to create consensus designed LRR proteins,[20] and some of these proteins have proved to be expressed and folded. A consensus design described for the tetratricopeptide repeat protein (TPR)[21] led to useful proteins with a high level of specificity and potential cellular applications. [22] , [23] and [24] Recently, highly expressed, stable and well folded Armadillo repeat proteins[25] have been designed with the aim to create peptide-dedicated scaffolds.

HEAT repeats were first identified in eukaryotic proteins but were later found as common motifs in prokaryotes as well.[26] The biological functions of these proteins are very diverse, but HEAT repeats are often involved in protein–protein interactions. Representative members of this family are PP65A and Importin β 1. These proteins are folded as a succession of α helix pairs forming a right-handed superhelix. The juxtaposition of helices results in extended surfaces recruited by evolution as target binding sites. In some cases the curvature and local flexibility of the elongated solenoid allow HEAT repeats to wrap around their protein partners. The HEAT repeat proteins have highly divergent repeat sequences and at least three different subgroups have been classified. This family is distantly related to the Armadillo repeat proteins. [26] and [27] Possibly because of this high variability, HEAT repeat motifs have not been explored for the design of artificial protein scaffolds.

The aim of the present study was to describe a new type of artificial repeat protein inspired by the analysis of HEAT repeat structures and sequences. Starting from a known structure of a thermostable HEAT repeat subfamily, we identified a well-defined consensus

sequence for stable HEAT repeats and devised adapted N- and C-caps. Based on this designed sequence, we built a test protein library and assayed the biophysical properties of proteins with a different number of repeats. In this library, the proteins differ by the number of repeats and by the local sequence of each repeat in variable positions. The tolerance to sequence variability and potentially functional positions was explored by characterization of artificial proteins selected randomly from this library.

Results

Design of α-repeat sequence

HEAT repeats are rather long (37 – 47 amino acids) and variable motifs compared to those from other repeat families. This highly divergent character could complicate the design of the consensus sequences required to make new artificial repeat proteins. Any sequence feature incorporated incorrectly into an artificial motif would be repeated in the resulting protein and could consequently have severe cumulative destabilizing effects. Therefore, if the consensus is defined from a widely divergent sequence family, the risk is high of incorporating features originating from several subfamilies that are no longer mutually compatible into a single consensus. We therefore focused on a specific class of HEAT repeat proteins classified as PBS HEAT-like repeat (SMART (SM00567) and Pfam (PF03130)), named from phycobylin synthase accessory protein, in which it was first identified. This repeat sub-type interested us for several reasons: first, these repeats are commonly found in thermophilic microorganisms and consequently could provide a stable protein platform. Second, compared to other HEAT repeat subgroups, this repeat family is shorter and its distribution appears to be more homogeneous in length and sequence. The 3D structure of a protein from this group, Mth187, has been solved (PDB code 1TE4, Fig.1b).[28] Mth187 is of unknown function and was originally selected as a target in a structural genomics project centered on the genome of the thermophile *Methanobacterium thermoautotrophicum* (Archaea).

In order to completely define the sequence features associated with this protein subgroup, sequence analysis was done, starting with the known structure of Mth187 as a guide. Mth187 is made by a succession of three pairs of α helices and a disordered N-terminal extremity. Each pair of αhelices corresponds to one repeat (30 or 31 residues/repeat). A Blast search was done against the sequence of the first two repeats of the folded part of Mth187.

The last repeat of Mth187 acts as a C-cap and was therefore excluded. Individual repeat sequences were extracted from the closest Blast matches and aligned to define a first consensus: GDERAVEPLIKALKDEDWYVRRAAAEALGEI

In order to further explore the protein sequence space around this consensus as well as the distribution of its sequence variability, a larger collection of closely related sequence motifs was then compiled. For the Blast search at this step, we used an idealized sequence made by five consecutive identical repeats corresponding to the previously established consensus:

$$(GDERAVEPLIKALKDEDWYVRRAAAEALGEI)_5$$

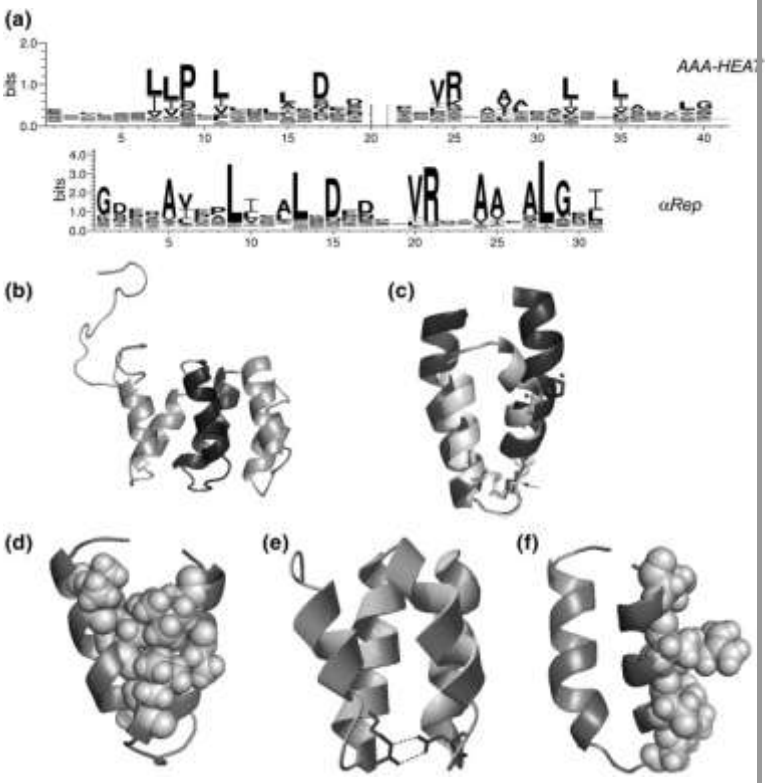


Fig.1:Design of the αRep motif. (a) A collection of protein sequences made of consecutive HEAT repeats related to Mth187 was compiled. The alignments of individual repeats are shown as a sequence logo using the sequence numbering in each motif family as the abscissa. Sequence conservation within the HEAT-PBS repeat subgroup (αRep) and comparison with the previously defined AAA-HEAT repeat subclass (AAA-HEAT). (b) The structure of the thermostable HEAT repeat protein used as a guide (Mth187, PDB code 1TE4). This protein is made of three consecutive HEAT repeats classified as HEAT-PBS repeats. The N-terminal segment of the chain is not folded. One HEAT repeat corresponds to two consecutives α-helices (dark gray). (c) The structural superposition of an AAA-HEAT repeat and an αRep HEAT repeat. Residues 396–434 from the PR65/A subunit of human protein phosphatase 2A (dark gray, PDB code 1B3U)[26] and residues 109–139 from αRep-*n*4-a protein (light gray, PDB code 3LTJ, this work) were overlaid. Positions *D15*, *V20* and *R21* in the αRep motif are superimposable on positions *D17*, *V24* and *R25* in the AAA-HEAT motif, as indicated by the arrow. The conserved proline P9 in the AAA-HEAT motif is not structurally equivalent to P8 in the αRep motif; these two proline side chains are shown (□). The two AAA-HEAT repeat helices are one helical turn longer than the αRep repeat helices. (d) Residues of the central Mth187 repeat corresponding to conserved apolar residues *A5*, *L9*, *L13*, *V20*, *A24*, *A25* and *L28* are shown as spheres. These conserved residues are associated with the packing of a single repeat between the two helices and the packing with neighboring repeats. (e) Conserved H-bonds between *R21* of one repeat and *D15* of the next repeat in the αRep context. (f) The highly variable positions *18*, *19*, *22*, *23*, *26* and *30* are mapped on the central repeat of Mth187 and shown as spheres. These hypervariable positions correspond to the external surface of helix 2.

This search procedure efficiently identifies natural sequences made from consecutive repeats closely related to the consensus and, presumably, to the Mth187 repeat structure. A set of 500 distinct repeats from the 100 closest matches to the penta-consensus was extracted and aligned. The sequence features of this aligned repeats ensemble are shown in Fig.1a. The different positions shared by the repeated motifs in the sequence collection are clearly not equally conserved (Table1). Some positions are highly conserved and others are hypervariable.

Table 1. Alignment of the original Mth187 repeats with the αRep randomized motif

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Mth187-a[a]	G	D	E	-	A	F	E	P	L	L	E	S	L	S	N	E	D	W	R	I	R	G	A	A	A	W	I	I	G	N	F
Mth187-b[a]	Q	D	E	R	A	V	E	P	L	I	K	L	L	E	D	D	S	G	F	I	R	S	G	A	A	R	S	L	Q	Q	I
Consensus 1[b]	G	D	E	R	A	V	E	P	L	I	K	A	L	K	D	E	D	W	Y	V	R	R	A	A	A	E	A	L	G	E	I
Randomized motif	G	D	E	R	A	V	E	P	L	I	K	A	L	K	D	E	D	X	X	V	R	X	X	A	A	X	A	L	G	X	I
Frequency[c]	0.56	0.34	0.26	0.23	0.71	0.40	0.34	0.24	0.87	0.42	0.14	0.47	0.81	0.16	0.75	0.3	0.46	-	-	0.75	0.86	-	-	0.78	0.63	-	0.69	0.89	0.71	-	0.48
Structure	Turn	h1	h1	h1	h1	h1	h1	h1	h1	h1	h1	h1	h1	h1	turn	turn	turn	h2	h2	h2	h2	h2	h2	h2	h2	h2	h2	h2	h2	h2	turn
C-cap (Mth187-c)[a,d]	G	G	E	R	V	R	A	A	M	E	K	L	A	E	T	G	T	G	F	A	R	K	V	A	V	N	Y	L	E	T	H
N-cap (A0B7C6)[e]	PLRA	D	P	E	K	V	E	M	Y	I	K	N	L	Q(K)	D	D	S	X(S)	X(N)	v	R	X(A)	A(Q)	A	A	X(E)	A	L	G	X(K)	I

[a] Mth187-a, -b and -c refer to the residues observed at this position in the first, second and third repeats in the Mth187 structure used as a guide (PDB Code: 1TE4). The sequence corresponding to the first two repeats Mth187-a, -b (from G20 to I80) was used for the initial blast search.

[b] Consensus 1 is the most common residue at each position in the multi-alignment of Mth187-like repeats resulting from the initial blast search.

[c] Frequency refers to the frequency of the consensus residue in an extended repeat database (see text).

[d] The last repeat of Mth187 (Mth187-c) was selected as a C-cap.

[e] The sequence selected as an N-cap (from protein A0B7C6) is indicated and aligned with the structurally equivalent position of the repeated motifs. The residues indicated in parentheses correspond to the residues found in A0B7C6. These were changed in position *14* to create a restriction site and in position *18, 19, 22, 23, 26, 30* to introduce variability in the N-cap.

When mapped onto the structure of Mth187 repeats, conserved positions can be subdivided in several groups. First, a set of apolar side chains *A5, L9, L13, V20, A24, A25, A27* and *L28*‡ contribute to the packing interaction between the two helices of the same repeat and to the stacking between neighboring layers (Fig.1d). Second, the highly conserved *R21* of each motif makes a buried electrostatic interaction with the carboxyl group of *D15* of the following motifs (Fig.1e). The conserved *G1* is part of the connecting turn between consecutive repeats and *G29* allows tight packing between helix 2 of consecutive repeats. The

remaining variable positions correspond mainly to the solvent-accessible positions of the two helices. As already noted for Mth187,[28] there is a clear asymmetry of the variability of these two helices. The outside positions of the first helix (*3, 4, 7 and 11*) are occupied by a high proportion of polar or helix-stabilizing side chains (K, R, E, Q and A). The most variables positions (*18, 19, 22, 23* and *26*) are located on the outside surface of the second helix of the motif and are occupied by a higher diversity of side chains and, notably, a significant fraction of aromatic side chains (Fig.1f). This suggests that the latter positions could be recruited for interaction with molecular partners. These positions were therefore selected as randomization sites in our design.

Comparison with known HEAT repeats sequences

The consensus defined here shares common residues (leucine *9, 13* and *28*) with other subclasses of HEAT repeats. [26] and [27] Among the three HEAT repeat subclasses defined earlier,[26] the present consensus appears to be closer to the AAA subclass, which comprises the majority of HEAT repeat proteins. The sequence alignment originally used to define the HEAT repeat AAA subclass[26] was compared to the ensemble of sequences used to define the αRep consensus using a logo representation (Fig.1a). The central part of our consensus (from *L9* to *A25*, Fig.1a αRep) is clearly similar to the central part of the AAA consensus (from *L11* to *A29* using AAA-HEAT sequence numbering; Fig.1a). The AAA-HEAT repeat family is, however, more divergent than the sequence ensemble used to define the αRep consensus. The AAA-HEAT repeats are also typically longer than αRep repeats and with additional residues usually located at the repeat extremities. Structural alignment of a prototypical HEAT repeat with an αRep repeat shows that the central parts of the repeats are similar with a remarkable superposition of the local conformation of some highly conserved residues (*D15, V20* and *R21*) (Fig.1c). Both helices of the AAA-HEAT family are extended by, on average, one helical turn relative to the αRep repeat. The first helix of an AAA-HEAT repeat is bent due to the highly conserved *P9* (AAA-HEAT numbering) residue, such as to maintain tight packing between the extremities of two helices. Although position *8* of our repeat family is also occupied by proline, this is not located in the equivalent position and is not so highly conserved. The packing of the hydrophobic side chains displays some differences between the two repeat families, but detailed comparison is complicated by the natural variability within the AAA-HEAT subclass. A recent update of sequence structure correlation in HEAT/ARM repeat proteins underlined the structural diversity of these natural repeats.[27]

Sequence of *N*- and *C-caps*

The external repeats at each end of most repeat proteins are referred to as N- and C-cap repeats. Capping repeats can have an overall fold similar to internal repeats but must differ at several sequence positions as one side of the cap, the exoface , will be exposed to the solvent, whereas the equivalent position in the internal repeats are part of the inter-repeat contacts.

The structure of Mth187 has one clear C-cap motif, folded as a pair of α helices. This sequence was therefore used directly, as a C-cap. The N-terminal part of the protein (residues 1–23) is not folded, and therefore cannot be used as an N-cap. We therefore searched HEAT repeat proteins for alternative N-cap candidates homologous to Mth187. Candidate N-cap sequences should appear as an N-terminal extension relative to a repeated consensus-like sequence. We retained a protein (UniProtKB identifier *A0B7C6* (***A0B7C6_METTP***)) with a very clear Mth187-like consensus preceded by a 34 amino acid residue N-terminal extension, whose length is close to what is expected for a two helix repeat (31 residues). Additionally, this putative N-cap sequence had characteristic features of the Mth187 HEAT-like repeat: LX DXXXXVRXXAAXALGXI suggesting this peptide forms an N-cap made by two helices. The positions of the putative N-cap second helix equivalent to hypervariable positions of the internal repeats are expected to be oriented towards the binding site and for this reason were also randomized. The sequences of the repeat motifs, of the N- and C-caps and of the natural protein used as an initial guide are summarized in Table1.

Library construction

Our primary objective was to evaluate whether the sequence design described above could give rise to folded proteins with favorable expression and biophysical properties. Our goal was to test the tolerance of the repeat sequences to variability in the potential binding sites positions as well as the relative stability of proteins with different number of repeats. These properties were explored using a test library made by a variable number of repeat modules contained between the N- and C-caps, each repeat having variable positions. For the library construction, we developed a new approach based on directional polymerization of a microgene corresponding exactly to one repeat, with variable positions encoded by degenerate codons. The process used to make this test library is illustrated by Fig.2.

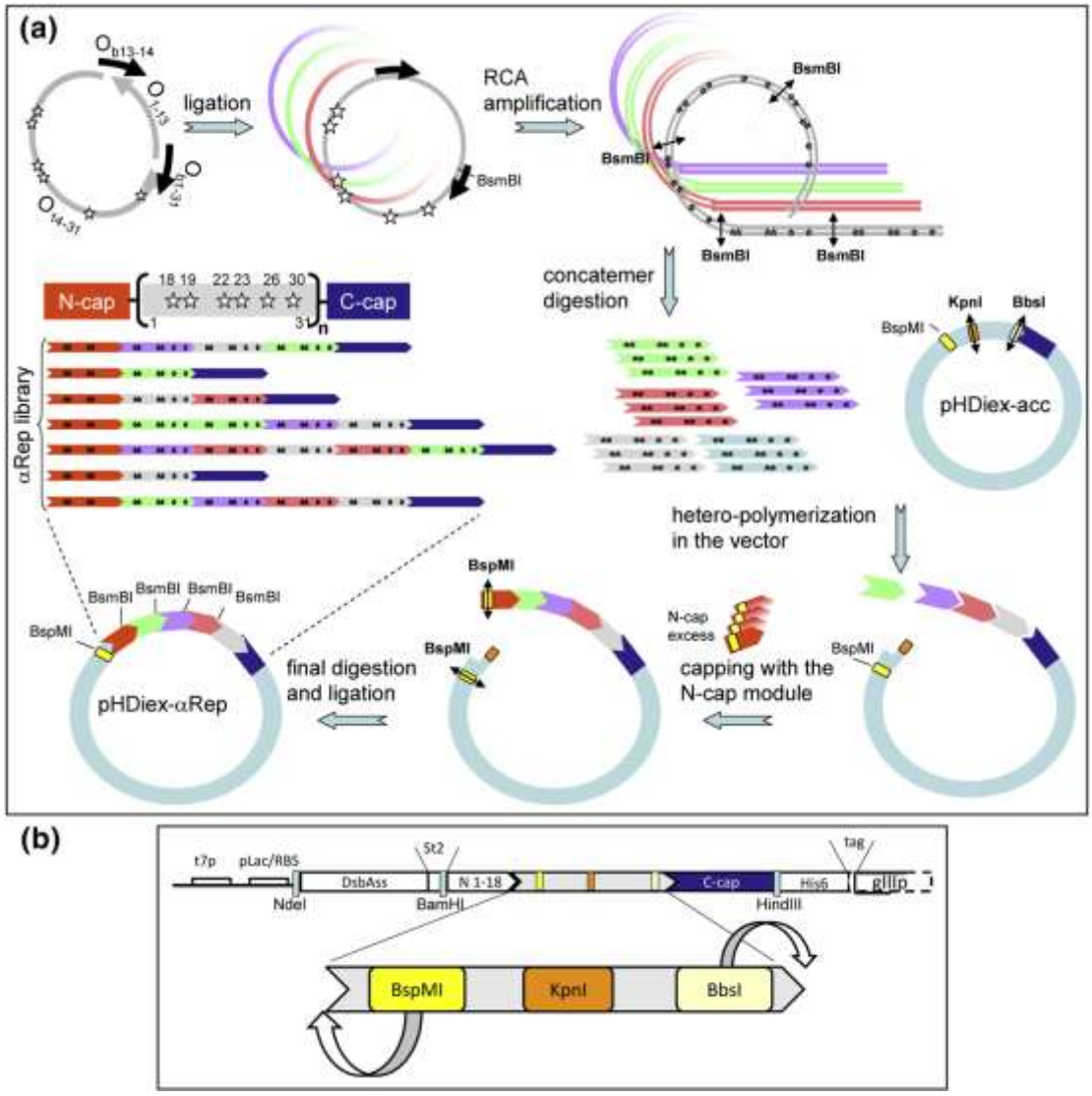


Fig.2:αRep library construction. (a) The 31 amino acid αRep motif was coded by two oligonucleotides including one degenerate 54-mer oligonucleotide (O_{14-31}) and a non-degenerate 39-mer oligonucleotide (O_{1-13}); degeneration is indicated by stars and corresponds to amino acid positions *18, 19, 22, 23, 26* and *30*. (Detailed oligonucleotide sequences are given in Supplementary Data Table S2). The oligonucleotides were annealed with the two short bridging primers O_{b13-14} (24-mer) and O_{b1-31} (27-mer) and ligated to form DNA circles coding for individual repeats. The circles were used as a matrix for RCA amplification to give double-stranded homopolymers of repeats. Digestion of the concatemer with a restriction enzyme (BsmBI) resulted in a collection of single repeats with different sequences. The repeats were heteropolymerized in the C-cap-containing acceptor vector (pHDiex-acc) digested at KpnI and BbsI restriction sites. Polymerization was capped by ligation with an excess of the N-cap variable module. The capped polymer was digested with BspMI and circularized by an intramolecular ligation to give pHDiex-αRep vectors. The resulting library encodes αRep proteins that differ in the motif sequence and the number of repeats. (b) A detailed view of the expression construct and acceptor vector design. The expression construct can be expressed from a t7 promoter (t7p) or Lac promoter depending on the bacterial strain used. RBS, ribosome-binding site. The coding sequence corresponds to the DsbA periplasmic signal sequence (DsbAss), StrepTag2 (st2), the constant part of the N-cap (N1–18), the repeated modules, the C-cap, and a His_6 tag fused through a suppressible stop codon (tag) to gene IIIp (gIIIp) of M13 (249–406). The acceptor vector (pHDiex-acc) used for heteropolymerization contains a synthetic cassette. BbsI cleavage of this vector creates sticky ends compatible with microgenes with BsmBI-cleaved repeats. The KpnI site is used as a protecting group to avoid reclosing the BbsI vector site during microgene polymerization. Once polymerized and capped by an excess of the variable part of the N-cap, the reclosing extremities are deprotected by BspMI cleavage and each molecule is recircularized by intramolecular ligation.

The repeat microgenes were synthesized as follows. Two oligonucleotides corresponding to the coding strand were annealed at their extremities with complementary bridging oligonucleotides to make DNA circles coding exactly for one repeat. The circles were closed by ligation and the pool of circularized repeats was amplified by rolling circle amplification using Phi 29 polymerase.[29] The resulting products are long homopolymers of repeats that were digested into monomers by restriction. A non-palindromic recognition site for the BsmBI enzyme is present in each repeat at positions coding for *G1-A2-G3* (/x xxx/ g**GA GAC GA** a). The monomeric repeats were then heteropolymerized by ligation onto the C-cap previously inserted in the destination vector (or acceptor vector, pHDiex-acc). The resulting heteropolymerized microgenes are capped by ligation with an excess of N-cap modules. Cohesive extremities for circularization were generated by cleavage of the N-capped polymers with the restriction enzyme BspMI. The vector was finally closed by an efficient intramolecular circularization step.

In this procedure, the number of modules in the library is variable and controlled by the respective concentrations of the acceptor vector and repeats at the heteropolymerization step.

Control of sequence variability in degenerated positions

In order to avoid potentially destabilizing or problematic amino acids such as proline within α helices, the variable codons were not fully randomized (e.g. by NNK codons). Instead, diversity was introduced using a set of partially degenerated codons (Table2). Position *23* was restricted to Ser or Ala, as a bias for small residues is observed in natural sequences at this position, presumably because this residue is oriented toward helix 2 of the preceding repeat and a large side chain might disturb the packing between repeats. Position *30* was restricted to E, K or Q because this codon partially overlaps the nucleotides corresponding to the sticky ends of the BsmBI site. These nucleotides must be kept constant for the microgene polymerization step. E, K and Q are among the commonest residues at this position in natural sequences homologous to Mth187. The randomization schemes of positions *18, 19, 22* and *26* were biased to eliminate residues with very low frequency and sample preferentially the commonest residues naturally observed in each variable position of the repeat. The sampled diversity was enriched specifically in Y and W because these residues tend to be favored in protein/protein interaction sites. It was essential, therefore, to test whether these residues are structurally acceptable in the potential binding site. However, the structure of the genetic code

necessarily limits the diversity encoded by a limited number of partially degenerated codons. For this reason, the sequence diversity is biased somewhat arbitrarily, but includes side chains with very different size, shape and polarity. This library is therefore well suited to explore the correctness and foldability of proteins resulting from this sequence design.

Table 2. Randomization scheme

Position[a]	Codons[a,b]	Encoded Amino Acids[c]
18	snd (4)	A, D, E, G, H, L, P, R, V
	tsg (1)	S, W
	tmy (1)	S, Y
19	khy (4)	A, D, F, S, V, Y
	van (1)	D, E, H, K, N, Q
	tgg (1)	W
22	vdr (5)	E, G, K, L ,M ,Q ,R ,V
	tsg (1)	S, W
23	kck	A, S
26	khy (5)	A, D, F, S, V, Y
	tgg (1)	W
30	vaa	E, K, Q

Table2. Randomization scheme

[a] For each position indicated, the degenerated codons used in the microgene synthesis and the corresponding encoded amino-acids are indicated.

[b] For positions 18, 19, 22, 26, the relative proportion of degenerated codons in the randomized oligonucleotides pool is indicated.

[c] Amino acid distribution computed from sequences of randomly picked clones in the experimental library corresponds to the encoded distribution

Characterization of the library

A library of 3×10^8 independent clones was constructed. The library construction procedure is efficient enough to make phage display libraries of useful size. The distribution of repeat numbers within the library can be visualized by a restriction of a plasmid pool from the library with restriction sites located on each side of the coding sequence (Fig.3). This gives a very clear pattern of discrete repeat polymers with a number of repeats between zero and 10, with a maximal frequency of two to five repeats.

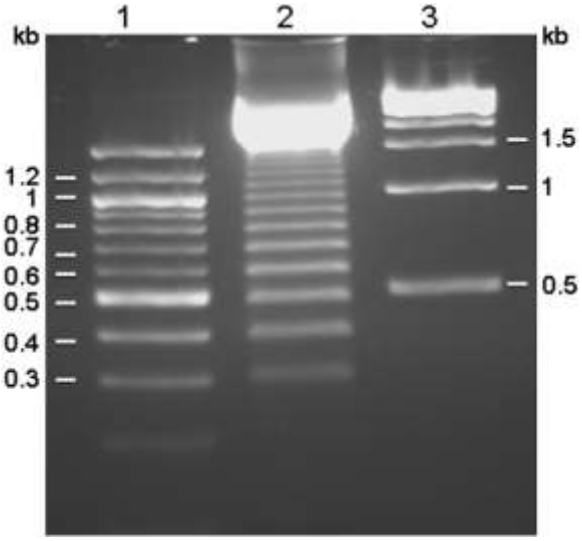


Fig. 3: DNA restriction of the αRep library. Lanes 1 and 3 correspond to 100 bp and 1 kb standard DNA ladders, respectively. A pool of plasmids from the pHDiex-αRep library was digested by restriction enzymes NdeI and HindIII located before the N-terminal and after the C-terminal extremities of the coding sequence, respectively. A sample of this digestion product was loaded in lane 2; 300 bp corresponds to the length of a protein- coding sequence with one repeat between the N and C-caps. Each additional repeat in the sequence corresponds to the addition of 93 bp in the DNA fragment.

The sequence of 49 randomly picked clones was determined. The sequenced clones had between zero and seven repeats, as expected from the length distribution observed by restriction on the plasmid pool. A set of 16 of these corresponded to expected sequences and the remaining 33 had at least one coding module with a frame-shift relative to the desired sequences. These out-of-frame clones could result either from errors during oligonucleotide synthesis, or from miss-assembly and amplification errors during the module production steps. Expected sequences were found for 77 of 112 individual sequenced repeats (68%), but the fraction of in-frame proteins decreases with the length (number of repeats) of these proteins. As a consequence, there is a low proportion of long coding sequences in this library.

Screening for soluble expression

In order to evaluate the proportion of expressed and soluble proteins in the library, a systematic test was done on a sample of 72 randomly selected individual clones. These clones were screened for soluble expression by a colony filtration blot (CoFi blot), which allows specific detection of the proteins expressed in a soluble form. [30] and [31] The CoFi blot experiment revealed that 33 % (24/72) of the sampled clones were expressed as soluble proteins (Fig.4a). The CoFi blot positive clones were further tested for expression at 37 °C in liquid culture and were shown to express soluble proteins (Fig.4b).

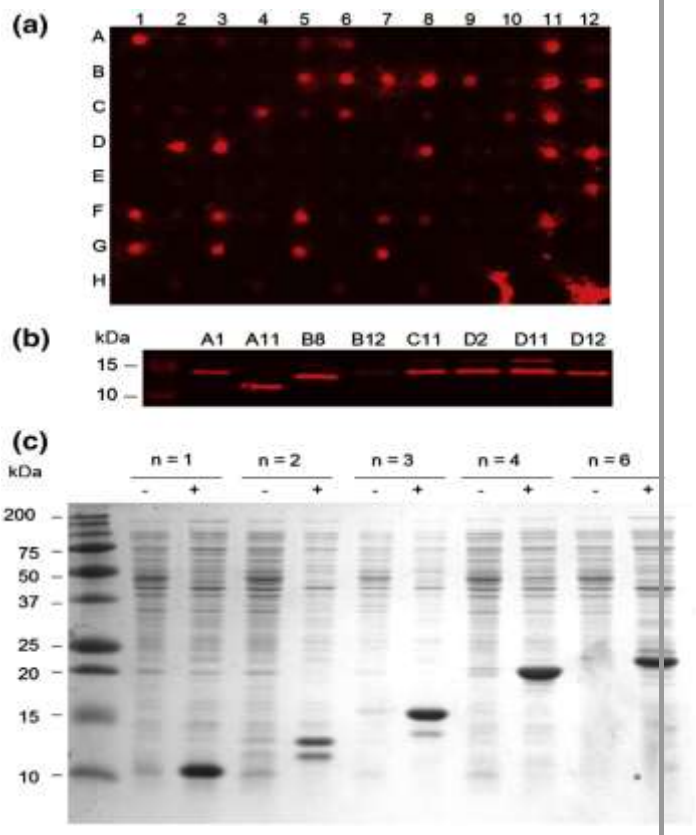


Fig.4. Expression properties of αRep proteins. (a) Screening for soluble expression by CoFi blot; 72 clones were chosen randomly from the pHDiex-αRep library and grown in a 96-well plate. Positive controls of well-expressed soluble proteins were grown in wells G1, G3, G5, G7, H10 and H12. Non-coding clones were used as negative controls in wells H2, H4, H6 and H8. (b) Western blot analysis of soluble expression from periplasmic expression vector (pHDiex) in liquid cultures. Eight positive clones from the CoFi blot experiment were expressed in liquid culture at 37 °C for 4 h after induction by IPTG. Soluble fractions obtained from bacterial lysates were analyzed by western blot using fluorescent immunodetection of the His tag. (c) Cytoplasmic expression of five αRep proteins (αRep-*ni*-a) with *i* = 1, 2, 3, 4 or 6 variable repeats. These proteins were screened as positive in CoFi blot and western blot experiments and their genes were sub-cloned in a cytoplasmic expression vector (pQE-31); liquid cultures were incubated at 37 °C with (+) or without (–)induction by IPTG. After bacterial lysis, soluble fractions were analysed by SDS-PAGE followed by staining with Coomassie brilliant blue. The expected molecular mass values calculated from the sequences are: 12.0 kDa (αRep-*n*1-a), 15.2 kDa (αRep-*n*2-a), 18.6 kDa (αRep-*n*3-a), 22.2 kDa (αRep-*n*4-a) and 28.5 kDa (αRep-*n*6-a).

These results suggest that (i) the proportion of clones in the library with a correct coding sequence is close to the fraction of expressed and soluble clones as detected by CoFi blot and (ii) the clones observed as positive in CoFi blots are, as expected, expressed efficiently as soluble proteins. Therefore, most of the non-expressing clones result from the fraction of the library corresponding to non-coding sequences (e.g. with one or more frame-shifted repeat). This was confirmed by sequencing non-expressing clones that appeared to correspond either to clones with out-of-frame inserts or to clones with no inserted repeat between the N- and C-caps. Taken together, these observations indicate that most in-frame sequences give rise to soluble proteins.

Cytoplasmic expression

The experiments described above were conducted with a combined phage display/expression vector used to construct the libraries. In these constructs, the protein are expressed as a fusion with a signal recognition particle-dependent DsbA signal sequence[32] and are secreted into the periplasm. The periplasmic expression level might be limited by the transport step and therefore might not reflect the expression yield using classical cytoplasmic expression constructs. We therefore monitored cytoplasmic expression efficiency for a subset of these proteins using the cytoplasmic expression vector pQE-31. Among the different clones tested for soluble expression, five with different numbers of repeats were chosen for further characterization. αRep-*n*1-a (*n* = 1), αRep-*n*2-a (*n* = 2), αRep-*n*3-a (*n* = 3), αRep-*n*4-a (*n* = 4) and αRep-*n*6-a (*n* = 6). The repeat sequences of these clones differ only, as expected, in the degenerate positions. Protein expression in liquid *Escherichia coli* cultures at 37 °C was induced by the addition of IPTG. The five proteins were over-expressed in a soluble form, as shown by SDS-PAGE (Fig.4c) and they could be purified with a yield of 10 – 90 mg of purified proteins per liter of shake-flask culture.

CD, stability and reversibility

The secondary structure content of the characterized proteins was analyzed by far-UV CD. For each αRep protein, the far-UV CD spectra were recorded at 25 °C under identical conditions (Fig.5a). All spectra are superimposable and display the characteristic signal of an all-α protein with a maximum absorbance at 192 nm and minima at 209 nm and 222 nm. A quantitative comparison showed that all characterized proteins have the same ellipticity on a per residues basis; this clearly suggests that the helicoidal content is correlated directly with

the length of the protein and that all the repeats of longer proteins are folded. Figure5b shows the CD spectra of the αRep-*n*3-a monitored at 25 °C, at 95 °C and after cooling to to 25 °C. At 95 °C, proteins lost most of their secondary structure. For all five αRep proteins (data not shown) the CD spectra recorded at 25 °C before and after denaturation are identical. These data indicate that thermal unfolding of secondary structure in these proteins is reversible.

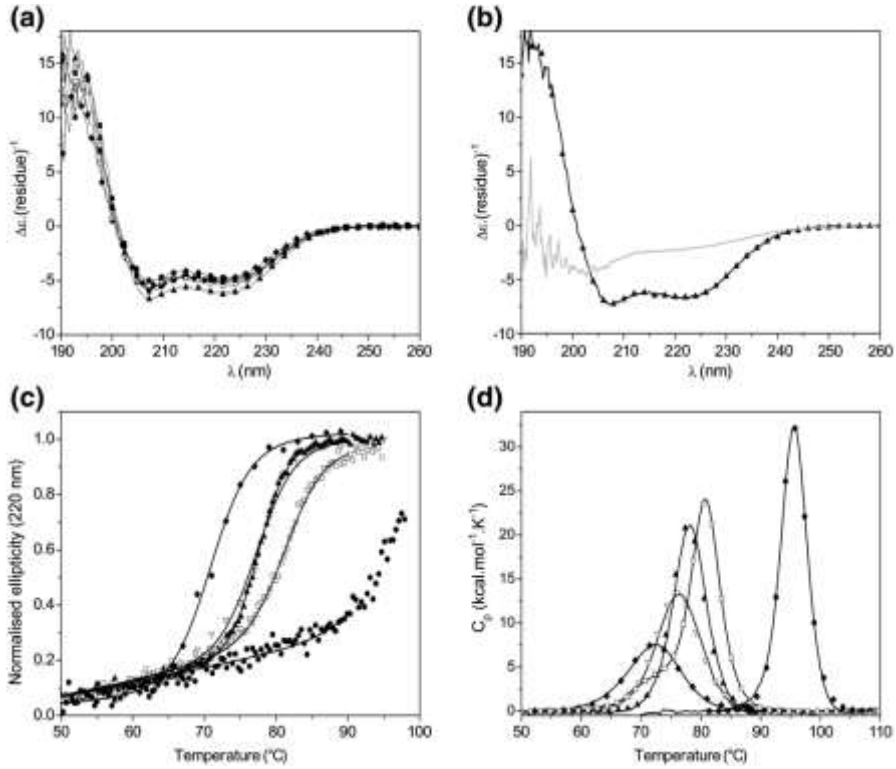


Fig.5. Biophysical properties of the αRep protein variants. (a) The CD spectra of αRep-*n*1-a (♦), αRep-*n*2-a (▽), αRep-*n*3-a (▲), αRep-*n*4-a (□) and αRep-*n*6-a (●). Each spectrum was measured with 10 µM protein. (b) The CD spectra of the αRep-*n*3-a (10 µM) measured at 25 °C (continuous black line), at 95 °C (continuous gray line) and after cooling to 25 °C (▲). This experiment was done for the other αRep proteins and presented spectra of similar shapes under these different temperature conditions (25 – 95 °C). (c) Unfolding transition curve assessed by the normalized variation of the ellipticity at 220 nm as a function of the temperature for the different αRep proteins (10 µM): αRep-*n*1-a (♦), αRep-*n*2-a (▽), αRep-*n*3-a (▲), αRep-*n*4-a (□) and αRep-n6-a (●). (d) Heat denaturation of αRep proteins assessed by DSC: αRep-*n*1-a (♦; 1.25 mg mL^{-1}, 104 µM), αRep-*n*2-a (▽; 0.23 mg mL^{-1}, 15 µM), αRep-*n*3-a (▲; 0.23 mg mL^{-1}, 12.3 µM), αRep-*n*4-a (□; 0.23 mg mL^{-1}, 10.3 µM) and αRep-*n*6-a (●; 0.23 mg mL^{-1}, 8.1 µM).

Thermal stability of αRep proteins

The thermal unfolding of all five proteins was first monitored by following the change in CD signal at 220 nm in the temperature range 35 – 95 °C (Fig.5c). The αRep-*n*1-a transition is sigmoidal with a midpoint T_m of 71.0(± 0.3) °C. αRep-*n*2-a and αRep-*n*3-a exhibit closely similar T_m values (77.4(± 0.5) °C and 77.6(± 0.1) °C, respectively; Table 3).

For αRep-*n*4-a, T_m was increased to 81.7(± 0.2) °C. The larger protein αRep-*n6*-a could not be thermally unfolded at all. This experiment showed that the stability of repeat proteins increased with the number of repeats.

Table 3. Analysis of heat denaturation monitored by CD and by DSC

	CD			DSC		
	ΔH_m *(kcal mol^{-1})*	Tm *(°C)*	ΔH_{cal} *(kcal mol^{-1})*	ΔH_{vH} *(kcal mol^{-1})*	Tm *(°C)*	$\Delta H_{cal}/\Delta H_{vH}$
αRep-n1-a	85.9 ± 4.9	71.0 ± 0.3	88.7 ± 0.4	80.7 ± 0.1	72.24 ± 0.03	1.1
αRep-n2-a	92 ± 7	77.4 ± 0.5	128.0 ± 6.0	101.0 ± 0.1	76.32 ± 0.03	1.27
αRep-n3-a	120 ± 1	77.7 ± 0.1	154.0 ± 0.4	134.0 ± 0.1	78.16 ± 0.05	1.3
αRep-n4-a	90.4 ± 1.5	81.7 ± 0.2	150.0 ± 6.5	153.0 ± 5.1	80.82 ± 0.07	0.98
αRep-n6-a	ND	> 95	186.0 ± 0.8	188 ± 1.0	95.55 ± 0.01	0.99

The thermal unfolding behavior of αRep proteins was also evaluated by differential scanning microcalorimetry (DSC) (Fig.5d). The unfolding of the αRep-*n*1-a protein resulted in a transition peak centered at 72.24(± 0.03) °C (Table 3). The $\Delta H_{cal}/\Delta H_{vH}$ ratio of 1.1 with a calorimetric enthalpy ΔH_{cal} of 88.7 kcal mol^{-1}, suggests a two-state transition. As for αRep-*n*1-a, the $\Delta H_{cal}/\Delta H_{vH}$ ratio of αRep-*n*2-a, *n*3-a, *n*4-a and *n*6-a were close to 1 (1.27, 1.3, 0.98 and 0.99, respectively). As observed by CD, the midpoint temperatures increased with the number of repeats with T_m values of 72.24 °C (*n*1), 76.32 °C (*n*2), 78.16 °C (*n*3), 80.82 °C (*n*4) and 95.55 °C (*n*6) (Table3). Except for the αRep-*n*4-a, the denaturation enthalpy increased significantly with the number of repeats: 88.7 kcal mol^{-1} (*n*1), 128.0 kcal mol^{-1} (*n*2), 154.0 kcal mol^{-1} (*n*3), 150.0 kcal mol^{-1} (*n*4) and 186.0 kcal mol^{-1} (*n*6).

Analytical size-exclusion chromatography (SEC)

Analytical SEC was used to compare the elution volumes of the αRep proteins (Fig.6). A solution of globular standard proteins was used as the control. First, we observed that none of the αRep proteins eluted at the expected volume for the standard calibration curve; they eluted at a rather lower volume; i.e. a higher Stokes radius. This is consistent with the non-globular form of these proteins. Their elongated form seems to confer on them the properties of larger proteins. Second, αRep-*n*4-a has a retention volume (10.33 mL) smaller than that of αRep-*n*6-a (10.67 mL), whereas this protein should appear with an elution volume around 12 mL, between the volumes observed for proteins with three and with six internal repeats. This

observation suggests that αRep-*n*4-a could be an exception within the αRep series and could have an oligomeric structure.

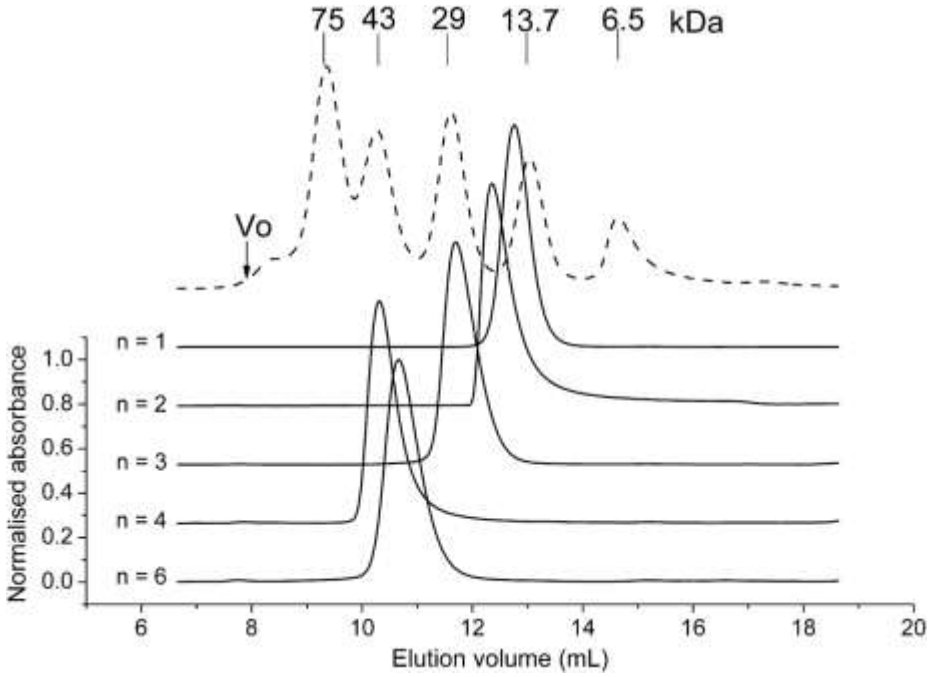


Fig.6. Elution profiles of αRep proteins obtained by analytical size-exclusion chromatography. Solutions of αRep proteins (100 µL, 3 mg mL^{-1}) with a variable number of repeats (*n* = 1, 2, 3, 4 or 6) were injected into an analytical Superdex 75 column equilibrated in 50 mM sodium phosphate, pH 7. For each elution profile, maximum $A_{280\,nm}$ was normalized to take into account the different tryptophan contents of each protein. The exclusion volume (V_0 = 7.9 mL), total volume (V_{tot} = 22.1 mL) and protein elution volume (V_e) were measured. The molecular mass of the αRep proteins calculated from their sequences was checked by mass spectroscopy: αRep-n1-a (V_e = 12.77 mL; 12.0 kDa), αRep-n2-a (V_e = 12.34 mL; 15.2 kDa), αRep-n3-a (V_e = 11.70 mL; 18.6 kDa), αRep-n4-a (V_e = 10.33 mL; 22.2 kDa) and αRep-n6-a (V_e = 10.67 mL; 28.5 kDa). A standard solution containing globular proteins was injected as a control (broken line). The protein standards used were: conalbumin (75 kDa), ovalbumin (43 kDa), carbonic anhydrase (29 kDa), ribonuclease A (13.7 kDa) and aprotinin (6.5 kDa).

Crystal structure

General organization

Protein αRep-*n*4-a was crystallized in two different forms (apo and PEG-bound) and its structure was solved to better than 2 Å resolution (Fig.7a; Table4). The crystal structure shows that αRep-*n*4-a associates as a dimer; this explains its smaller elution volume in SEC compared to αRep-*n*6-a. Each monomer is folded as a succession of α helix pairs stacking on each other. The succession of repeats forms a flattened right-handed superhelix. The structures of the internal repeats are almost identical, with the exception of the variable side chains. The relative positions of the repeats are also highly regular: helices 1 (or 2) of

successive repeats are parallel with each other and close to perpendicular to the long axis of the solenoid. The inter-helical distance between two consecutive repeats is shorter for the second helix (7.5 Å) than it is for the first helix (12.2 Å) of the repeat. This creates a curvature along the long axis of the super helix as observed for other α solenoids[33]. The convex face is made by helix 1 of successive repeats and the concave face of the protein is made by the succession of the external surfaces of helices 2. The variable side chains are clustered together and create a continuous surface located on the concave face of the protein (Fig.7b).

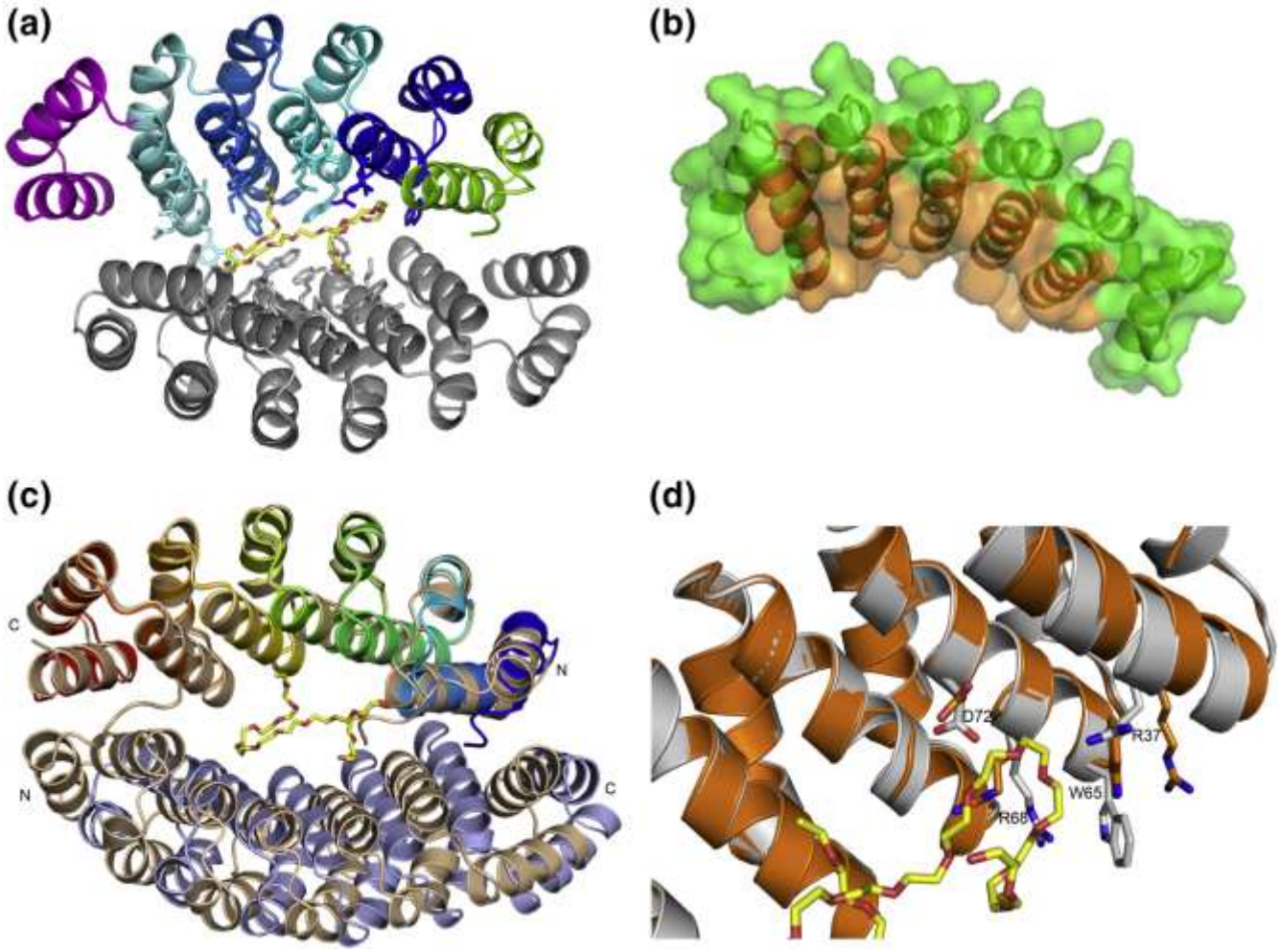


Fig.7. Crystal structure of the protein αRep-*n*4-a. (a) Structure of the PEG-bound dimer. The protein is shown in ribbon representation. The first monomer is colored gray. For the second monomer, the N-cap is colored green, the C-cap in purple and each internal repeat in blue. Side chains of the randomized residues are displayed as sticks representation and are located at the dimer interface. The bound PEG molecule found in the crystal structure is in yellow. (b) Molecular surface representation of a monomer. The surface formed by constant residues is indicated in green; the randomized residues form a continuous surface indicated in orange (N-cap and four internal repeats). (c) Structure superposition of the apo (brown) and the PEG-bound (blue) dimers. (d) Side chain conformational changes between the apo (gray) and the PEG-bound (orange) forms.

Table 4. X-Ray data collection and refinement statistics

				SeMet	
	Apo-form	**PEG-form**	**Peak**	**Inflexion**	**Remote**
Resolution (Å)	30-1.8 (1.85-1.8)	30-2.15 (2.21-2.15)	50-2.3 (2.36-2.3)	50-2.3 (2.36-2.3)	50-2.7 (2.31-2.7)
Wavelength (Å)	0.9791	0.933	0.979	0.9793	0.954
Space group	$P2_12_12_1$	$P2_1$		$P2_1$	
Cell parameters					
a (Å)	34.2	42.4		42.2	
b (Å)	64.6	52.4		52.6	
c (Å)	85.2	84		83.9	
β (°)		= 91.4		91.8	
Total number of reflections	96,558	75,520	102,415	103,658	42,217
Total number of unique reflections	17,826	20,242	29,713	29,876	19,433
R_{sym} (%)[a]	6.4 (42.6)	7.9 (44.6)	4.8 (11.2)	6.3 (26.6)	5.7 (29.5)
Completeness (%)	97.8 (97.6)	99.7 (99.5)	92.6 (62.3)	93.1 (63.2)	98.5 (99.2)
I/σ(I)	17 (2.9)	14.4 (3.2)	16.9 (7)	15 (3.6)	11.1 (3.1)
Redundancy	5.4	3.7	3.4	3.4	2.2
Refinement					
Resolution (Å)	30-1.8	30-2.15			
R/R_{free} (%)[b]	17.4 / 22.7	17.4 / 25			
r.m.s.d. bonds (Å)	0.005	0.007			
r.m.s.d. angles (°)	0.982	1.05			
Ramachandran plot					
Most favoured (%)	99.4	98.8			
Allowed (%)	0.6	1.2			
PDB code	3LTJ	3LTM			

Values in parentheses are for highest resolution shell.

[a] $R_{sym} = \Sigma_h\Sigma_i|I_{hi} - \langle I_h\rangle|/\Sigma_h\Sigma_i I_{hi}$, were I_{hi} is the *i*th observation of the reflection h, while $\langle I_h\rangle$ is the mean intensity of reflection h.

[b] $R_{factor} = \Sigma\ ||F_o| - |F_c|| / |F_o|$. R_{free} was calculated with a small fraction (5%) of randomly selected reflexions.

The N-cap

The structure of the segment chosen as the N-cap in our design was not known. It is folded as a pair of α helices, with an overall N-cap/repeat 1 arrangement very close to the inter-repeat contacts and efficiently shields the inter-repeat interface from the solvent on the N

terminus of the monomer. This is essentially due to three positions in the first helix of the consensus sequence substituted, in the N-cap repeat, by polar side chains: *A5*, *L9* and *A12* were changed to *K5*, *Y9* and *N12*, respectively. The randomized positions of the N-cap are structurally equivalent to the randomized positions of the internal repeats. The N-cap module is packed on the following repeat with the same relative orientation as the relative orientation of consecutive internal repeats.

The C-cap

The structure of the C-terminal part (repeats 3, 4 and C-cap) of the protein was compared to that of the guide structure used for the library design (Mth187 structure, PDB code 1TE4; rmsd of 1.9 Å over 85 Cα atoms). In these two structures, the packing of the C-cap differs from the packing between internal repeats. The angle between the second helix of the penultimate repeat and the first helix of the C-cap is more open than in preceding repeats. This leaves a space between the C-cap and the preceding repeat into which the second helix of the C-cap is partially inserted.

The dimer

As stated above, we have solved the structure of αRep-*n*4-a in two different space groups resulting in an apo structure (crystals grown in the presence of 2-methyl-2,4-pentanediol (MPD) as a precipitant) and a PEG-bound structure (crystals grown in the presence of PEG 1000 a as precipitant). Analysis of this protein by SEC indicated strongly that it forms dimers in solution. In both crystal forms, a large contact area is observed between similar portions of neighboring molecules. In the dimer, the N-cap from one monomer contacts the C-cap from the second monomer and *vice versa*. As a result, the concave faces of each monomer face each other and the long axis of the solenoids are almost anti-parallel, creating a deep crevice. The dimers do not superimpose well due to a 5 Å translation along the longitudinal axis of one monomer relative to the other (Fig.7c). This discrepancy is very likely to result from the presence in one crystal form of a PEG molecule bound in the deep crevice formed at the dimer interface. Gel-filtration chromatography and small-angle X-ray scattering (SAXS) experiments show that the dimeric structure of this particular variant is stable in solution, even in the absence of the bound PEG molecule. In the dimer of the PEG-bound form, both the PEG-binding residues and the monomer/monomer interface originate mainly from the randomized side chains. More specifically, Y33, Y34 and

R37 from the N-cap, W65, from internal repeat 1, W96 from internal repeat 2, W126 and F127 from internal repeat 3, and W158 and Q161 from internal repeat 4 are directly involved in the interface.

Interestingly, comparison of the structures in the apo and PEG-bound forms reveals two major differences. First, due to the change in orientation between the two monomers, the cavity appears wider in the PEG-bound form. Second, this relative displacement of the monomers is related to side chain rearrangements. The side chains from residues R37 of the N-cap and W65, R68 and D72 of internal repeat 1 that are contacting the PEG molecule adopt radically different conformations between the apo and PEG-bound structures (Fig.7d). Altogether, these observations suggest some flexibility within this scaffold.

Small-angle X-ray scattering

The SAXS data were recorded for all five αRep proteins at different concentrations. The oligomerization state of proteins in solution is derived from the value of the intensity at the origin $I(0)$, which is directly proportional to the molecular mass of the scattering object. The calculated scattering patterns of the atomic models built on the basis of the αRep-*n*4-a crystal structure (see Materials and Methods) were compared to the experimental curves.

All proteins are monomeric at low concentrations (< 10 µM; Fig.8a–c, and f), except αRep-*n*4-a, which is present in solution as a dimer (Fig.8e). The agreement between experimental curves extrapolated to infinite dilution and calculated curves of all protein models is excellent.

For αRep-*n*4-a, the stable state in solution is the dimer. The scattering pattern exhibits no concentration dependence and is perfectly described by the curve calculated using both dimer crystal structures (Fig.8e).

At higher concentrations (> 100 µM), proteins αRep-*n*1-a and αRep-*n*2-a remained monomeric whereas clear oligomerization patterns were observed for αRep-*n*3-a and αRep-*n*6-a.

In the case of concentrated solutions of αRep-*n*1-a and αRep-*n*2-a (> 2 g L^{-1}, i.e. 166 µM and 131 µM, respectively), scattering patterns exhibit weak attractive interactions but clearly no oligomerization

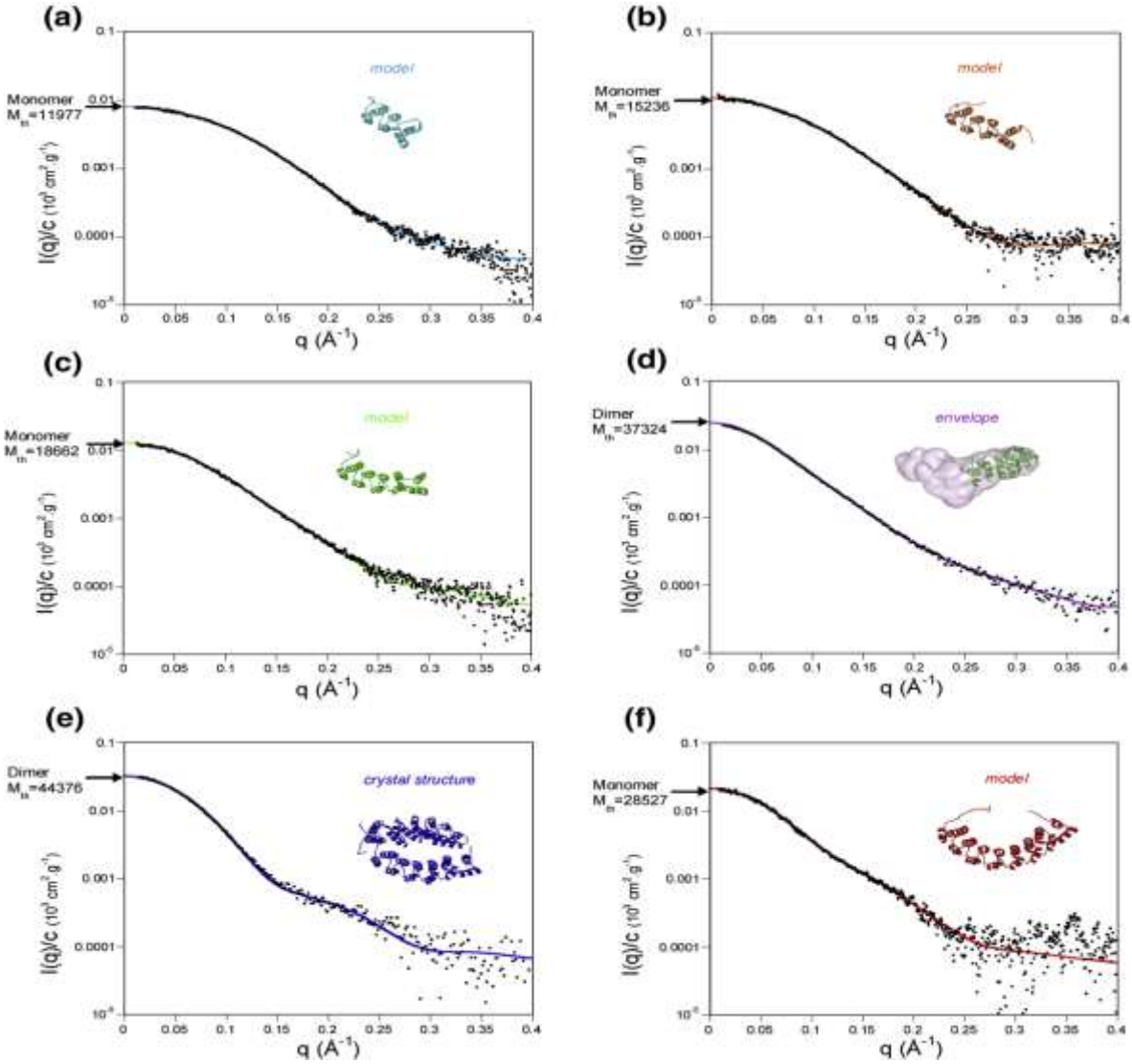


Fig.8. Analysis of αRep proteins by SAXS. Experimental scattering curves (dots) obtained by extrapolation to infinite dilution of curves of (a) αRep-*n*1-a, (b) αRep-*n*2-a, (c) αRep-*n*3-a, (e) αRep-*n*4-a and (f) αRep-*n*6-a. The theoretical value of the intensity $I(0)$ (in cm^{-1}) calculated from the sequence-derived molecular mass is reported on the *y*-axis in a–c, e and f. These values show that at low concentrations (10–50 μM) all proteins are monomeric with the exception of αRep-*n*4-a, which is dimeric. (d) Dimeric αRep-L1-n3 at a high concentration (10.1 g L^{-1}, 541 μM). Continuous lines correspond to calculated scattering patterns of models a – c and f or of (e) the crystal structure. The continuous line in d is the calculated curve of the model shown in the panel obtained with GASBOR for a chain of 2 × 170 residues. A monomer of αRep-*n*3-a has been positioned within this volume to help visualize the relative sizes.

In the case of αRep-*n*3-a, attractive interactions cause protein dimerization. (The protein concentration dependence of the radius of gyration and of the average molecular mass of scattering objects is shown in Supplementary Data Fig. S1A). A clear transition is visible with a midpoint close to 2 g L^{-1} (107 μM); at concentrations higher than 4 g L^{-1} (214 μM) the radius of gyration is practically constant and the (constant) mass of scattering objects is compatible with that of a dimer. The next step of data analysis involved simulated scattering patterns computed differently for the monomer (Fig.8c) and the dimer (Fig.8d). Regarding the monomer, the calculated scattering pattern of the atomic model was fit against experimental

data extrapolated to infinite dilution. Figure8c shows the structural model used and the calculated pattern, which fits the experimental data perfectly. In the case of the dimer, we first built a structural model with two monomers facing one another as observed in the crystal structure of αRep-*n*4-a. Such an arrangement cannot account for the experimental data. We then determined the envelope of the dimer using the *ab initio* program GASBOR,[34] which describes the scattering object as a chain of (2 × 170) dummy residues. The resulting envelope is elongated and large enough to accommodate two monomers in a tandem arrangement (Fig.8d). The chain of a monomer is superimposed on the envelope. The second monomer is not shown because nothing is known about the region of interaction. Finally, although the detailed structure of this weakly stable dimer is unknown, these results show clearly that the interaction between two αRep-*n*3-a monomers is different from that between two αRep-*n*4-a monomers.

Finally, αRep-*n*6-a exhibits an even stronger oligomerization propensity. At a low concentration (0.15 g L^{-1}, 5.2 μM) the protein is monomeric (Fig.8f) but the average molecular mass of scattering objects increases strongly with concentration and is larger than that of a dimer at a concentration of 1 g L^{-1} (35 μM). The highest concentration investigated was not high enough to determine whether the protein associates at high concentration into a well-defined oligomer (Supplementary Data Fig S1B).

Discussion

Sequence definition of the repeated motif

The sequence of αRep proteins was defined using a simple procedure based on consensus design. Throughout the procedure, care was taken to incorporate into alignment only those HEAT sequences closely related to the guide structure Mth187. Despite this relatively restrictive search procedure a large collection of repeat proteins was identified, allowing a precisely defined consensus. A high level of variability of side chains was found in some specific positions of this consensus. Those positions are predicted to be clustered on the same face of the folded proteins and to be involved in binding interactions. The consensus αRep motif is included in previously established consensus for HEAT repeats subfamilies[26] or to profiles available in databases for this subgroup. However, the procedures used here lead to a more precisely defined sequence family with respect to repeat length and diversity.

Our experience, as well as the literature, suggests that in artificial repeated proteins, the consequences of imperfections in repeat sequence design are easily amplified by the repetitive structure of the proteins up to the point where the resulting proteins are no longer able to fold. Therefore, although many more distantly related HEAT repeats could presumably be detected using more generalized consensus profiles or hidden Markov models, our objective was not to detect in a comprehensive manner all distantly related HEAT repeats in known genome sequences, rather to identify with high confidence a sequence motif that could give rise to a folded and stable protein and to explore the acceptable sequence variability allowed within these mutually compatible sequence repeats.

HEAT repeats have been described as distantly related to Armadillo repeat proteins; [26] and [27] however, this sequence is clearly distinct from the consensus sequences described for designed Armadillo repeat proteins.[25]

It has been suggested that covariances are important features of protein evolutive families. [25] , [35] and [36] In this case, the αRep consensus was established from a collection of closely related sequences and a simple consensus directly gave rise to folded proteins and thus analysis of co-variations was unnecessary. In more divergent sequence families, covariations could well be critical features to include in the repeat design.

Synthesis procedures

The procedure used to construct the protein library differs in several aspects from the methods described earlier.[15]. First, the gene segments coding for individual repeats were synthesized by circular amplification and cleavage of the resulting homopolymers. This method was retained because it is a convenient way to produce large amounts of highly polymerizable repeat sequences. In our hands, the conventional synthesis procedures based on PCR produce monomeric repeats fragments that often are incompletely cleaved in subsequent restriction. An incorrect extremity acts later as a polymerization terminator at the polymerization step resulting in low efficiency construction process. Although a range of technical solutions could probably be used to overcome this problem, circular amplification appeared to be an efficient way to generate large pools of degenerated highly polymerizable fragments. Rolling circle amplification (RCA) is commonly used for genome amplification and is known to preserve the diversity of complex sequence collections. RCA has been

suggested as an efficient way to amplify diversity in phage display selection experiments. [37] and [38]

The micro-genes were then directly polymerized in the vector, rather than assembled step-by-step. This approach was introduced in order to avoid PCR amplification, as PCR amplification of repeated sequences, although possible, can be problematic due to internal sequence repetition. In-vector-repeats polymerization results in libraries with a distribution in length for the coded proteins.

In this library, the sequence diversity in variable positions was evaluated experimentally by comparison of the sequences of 80 independent repeats and corresponded to the diversity expected from the coding scheme, with no clear diversity bias. Future improvements of this library construction procedure will be focused on minimizing sequence errors to maximize the fraction of long coding sequences and in a better sampling of sequence diversity in variable positions.

Biophysical properties

Soluble expression tests (CoFi blot and western blot) of randomly selected clones in the resulting library indicated that most clones with a coding sequence containing at least one insert were expressed as soluble proteins. Non-expressing clones resulted from either non-coding sequences due to incorrectly synthesized sequences or to coding sequences with no repeat inserted between the N- and C-caps. The subset of proteins (αRep-*n*1-a – αRep-*n*6-a) tested for cytoplasmic expression was efficiently expressed and purified. The thermal stability of the proteins was further characterized by CD and DSC. The proteins show a sharp thermal transition with T_m values ranging from 70 °C (one internal repeat) to more than 90 °C (six internal repeats). This shows that although their sequence is repeated, the proteins are folded as a single cooperative unit, similar to single-domain native proteins, and not as a molten globule or as a collection of independent repeats folding units. As observed for other repeat families, the overall trend is that protein stability increases with the number of repeats, although the sequence of each repeat and the combination of variable position can modulate this general trend. For example, a second four repeats αRep was recently purified and characterized by DSC. The stability of this protein (T_m = 87.7 °C by DSC) differs significantly from the stability of the four-repeats protein αRep-*n*4-a (T_m = 80.8 °C) described above.

Beside a direct sequence effect on stability, the difference in oligomeric structure between proteins could contribute to the modulation of stability between the different variants of the same length.

Oligomeric structure

Based on gel-filtration and SAXS experiments in solution and the crystal structure of αRep-*n*4-a, the proteins isolated from the library were either monomeric or dimeric (for αRep-*n*4-a) at low micromolar concentrations. For high micromolar concentrations, an equilibrium between monomers and dimers was observed for αRep-*n*3-a and oligomers larger than dimers seem to form for αRep-*n*6-a at high concentrations. No evidence for aggregation was observed for any of these proteins. Our data support the fact that different modes of oligomerization can be observed: while αRep-*n*4-a is an authentic dimeric protein, the oligomeric structures of the other proteins are only weakly stable. Furthermore, the SAXS results indicate that the dimeric form observed by SAXS for αRep-*n*3-a at high concentrations must have a more elongated shape than the symmetric shape observed for αRep-*n*4-a. Finally, in the crystallized dimeric structure, a subunit interface results from the N- and C-caps and from randomized side chains and therefore the stable dimerization is due to the sequence of this specific variant. The tendency to oligomerization might be enhanced by the randomization scheme used in this library, which introduces a high proportion of aromatic side chains in variable positions.

Mth187, the natural protein used as a guide to design αRep, was shown to be in equilibrium between a monomer and a dimer. The monomeric structure of Mth187 observed by NMR[28] was reported to be related to the presence of detergent (Chaps) used in the NMR sample preparation. It has been suggested that the dimeric interface could be located on the side of the protein corresponding to hypervariable side chains, exactly as observed here in the dimeric structure of αRep. Future experiments will attempt to identify the sequence determinants of αRep required for dimerization.

Applications

αRep proteins form an attractive scaffold for molecular recognition. The proteins are stable, well expressed and secreted efficiently. The large molecular surface can accommodate a diverse range of side chain combinations. Efficient methods to create libraries with variable

numbers of repeats have been developed. Selection methods such as phage or ribosome display will be used to select specific binders against protein targets. Additionally, the crystal structure of a dimeric variant has a deep crevice between two monomers making a fortuitous PEG-binding site. The comparison of the apo and the PEG-bound structures indicates some structure flexibility at the interface between monomers. Therefore, if sequence features governing dimerization could be established and controlled, αRep could be a well-adapted scaffold for protein recognition and for host–guest chemical studies and molecular recognition of small molecules.

Materials and Methods

Sequence analysis

The sequence of repeats 1 and 2 of protein Mth187 (G20 – I80) was used for a Blast search using a non redundant sequence database (Uniref 50). The aligned parts of the closest Blast matches (81 sequences with $E < 0.01$) were extracted, manually split into repeats and aligned. The commonest residues in this sequence collection, at each position of the repeat, were used to define consensus 1. An extended collection of repeat sequences was then established using a Blast search in Uniref 50, with a sequence made of five consecutive consensus 1 sequences as the search sequence. The 100 best "penta-consensus 1" similar sequences were collected ($E < 10^{-17}$) (Supplementary Data Table S1). The repeat sequences matching the probe sequences were extracted, split into their sequence modules and aligned. The sequence logo[39] used to visualize sequence conservation (Fig.1a) was created using the WEBLOGO server§. The AAA-HEAT repeat logo was computed from the multiple sequence alignment originally used to define the AAA-HEAT profile.[26]

Although this collection of repeats was created using similarity to a fully defined penta-consensus 1, the sequence variability in the aligned repeats is not uniform, it is much higher in some positions than in others (Fig.1a and Table1). Finally, in order to better appreciate the sequence variability at each variable position (*18, 19, 22, 23, 26* and *30*) without bias toward the residues occupying these position in consensus 1, a new collection of repeats was created using as a Blast search sequence an idealized sequence made of 10 consecutive repeats in which the variable positions were not specified.

Presumably because of the coherence of the sequence ensemble in this protein repeat subgroup, the process used to define the consensus is relatively robust. Minor variations in the procedure used to construct the repeat sequences alignments such as number of repeats in the search sequence, local sequences variations of the consensus, or threshold used to establish the repeat collection retained, lead essentially to the same final consensus and variability (not shown).

Expression vectors

Two different vectors were used. As the final usage of a library will be phage display selection of specific binders, a combined phage display and expression vector was used for library construction and periplasmic expression (pHDiex).[9] For this work, the construct described earlier was slightly modified: the signal sequence used for periplasmic expression was replaced by a signal recognition particle-dependent signal sequence (DsbAss), which has been described as more efficient than SEC-dependent sequences to display very stable proteins.[40] For cytoplasmic expression, a restriction fragment (BamHI/HindIII) containing only the coding sequence from the mature αRep proteins, without tags or export sequences, was subcloned into the pQE-31 expression vector. In this construct, the protein is expressed with a His_6 tag fused to the N-terminal extremity.

Synthesis of the microgenes

Two 5′-phosphorylated oligonucleotides O_{1-13} and O_{14-31} corresponding to the coding strand of the repeat sequences were hybridized with two "bridging" oligonucleotides, O_{b13-14} and O_{b14-31}, complementary to the extremities of the coding strand oligonucleotides. The hybridized oligonucleotide mix (5 µM each) was ligated by T4 ligase to give circular products that were used as substrate for RCA with Phi29 polymerase (TempliPhi kit, GE Healthcare). A 5 pmol (1 µL) sample of circularized product was mixed for 15 h at 30 °C in 20 µl of amplification reaction mix. The polymerized product was incubated at 65 °C for 15 min to inactivate the polymerase, diluted to 200 µl with water and appropriate buffer and restricted (20 U of BsmBI) for 4 h at 55 °C. At this stage, agarose gel electrophoresis of the cleaved amplified products showed a clear 100 bp band as expected from the length of the amplified sequence. The same treatment was done in parallel for the six oligonucleotides (O_{14-31}-a – O_{14-31}-f) corresponding to different combinations of the partially degenerated codons in the variable part of the repeat. Once amplified, the cleaved products corresponding to these

different sequences were mixed in stoichiometric amounts and co-purified using a NucleospinDNA purification kit (Macherey-Nagel).

Library construction

The library was constructed by polymerization of synthetic microgenes corresponding to repeats in a phage display/periplasmic expression vector (Fig.2a). An intermediate construct termed acceptor vector (pHDiex-acc) was first constructed (Fig.2b) and a synthetic cassette encoding three restriction sites (BspMI, KpnI and BbsI) was inserted between the N-terminal constant part of the N-cap (residues 1 – 18) on one side and the C-cap on the other side. This acceptor vector was first cleaved by KpnI and BbsI, which generated a single repeat-compatible cohesive end on the C-cap side. For the library construction, 19 µg of prepared vector was ligated with 3.2 µg of repeat microgenes in 300 µl (vector/repeat molar ratio approximately 1:6), for 15 h at 15 °C. The reaction was capped with an excess of N-cap variable part sequence (residues 19–35) as follows: 50 pmol of hybridized oligonucleotides N-stop and N-stop$_{rev}$ were hybridized, then added and ligated to the repeat/vector polymerization mix in a final volume of 500 µl. The ligated product was purified on Nucleospin columns and cleaved by BspMI. At this stage the BspMI site is present on the N-capping motif and on the constant part of the N-cap at the other end of the vector. Cleavage by BspMI generates linear constructs with cohesive ends that can be closed efficiently by intramolecular ligation. The products obtained following this final ligation were purified/desalted on Nucleospin column and eluted in 50 µl of water. DNA was electroporated in XL1-Blue MRF' electrocompetent cells using standard conditions. Eight 100 µl samples of electrocompetent cells were electroporated, and plated on 20 cm × 20 cm plates of agar in 2YT medium containing 200 µg mL^{-1} ampicillin and 1% (w/v) glucose. Appropriate dilutions of the electroporated cells were plated separately to evaluate the size of the library. Colonies were harvested after overnight growth, pooled and stored at –80 °C in 2YT medium, 20 % (v/v) glycerol. A sample of pooled bacterial cells was used to prepare a DNA pool for restriction analysis (Fig.3) and collective transformation in a high-level expression host.

Monitoring for soluble expression

The αRep library was transformed in the *E. coli* strain BL21(DE3)pLysS. Individual colonies were picked randomly and grown at 37 °C overnight in a 96-well microplate in 100 µL of 2YT medium containing 200 µg mL^{-1} ampicillin and 1% glucose. Glycerol was

added to each well and the plate was stored at –80 °C. This master plate was used as a matrix for the expression screening experiment. A CoFi blot experiment was done as described to detect proteins expressed in a soluble form. [10] and [30] The master plate was replicated on a 2YT agar plate containing 200 µg mL^{-1} ampicillin and 1% glucose. After an incubation overnight at 37 °C, the colonies were transferred onto a 0.45 µm pore size Durapore filter (Millipore) and expression was induced by IPTG at 37 °C for 4 h. The filter was then treated as described.[30] Soluble proteins were recovered on a nitrocellulose membrane and immunodetected with a mouse anti-His antibody (Qiagen) followed by an Alexa Fluor® 680 labeled anti-mouse antibody (Molecular Probes). Fluorescence detection was done with an Odyssey® Infrared Imaging System (Li-Cor) with excitation at 680 nm and emission at 700 nm. Positive clones were detected by quantitative fluorescence signal analysis.

Protein expression and purification

The gene coding for each αRep protein was subcloned into the pQE-31 vector (Qiagen). The corresponding plasmid was transformed into the expression *E. coli* strain M15[pREP4] (Qiagen). Cells were grown at 37 °C in 2YT medium containing 200 mg L^{-1} ampicillin and 25 mg L^{-1} kanamycin to an absorbance of 0.6 at 600 nm. Protein expression was induced by addition of IPTG to 1 mM and the cells were further incubated for 4 h. The cells were harvested, suspended in TBS, submitted to three freezing/thawing cycles and treated with lysozyme and benzonase for 30 min. After centrifugation, the His-tagged proteins were purified from the supernatant using nickel-affinity chromatography (Ni-NTA agarose, Qiagen) followed by SEC (HiLoad™ 16/60 Superdex™ 75 preparation grade column, GE Healthcare). For each protein, the purity of the final sample was checked by SDS–PAGE with an overloaded gel showing one well-resolved band with no visible contamination.

Circular dichroism measurements

CD spectra were recorded from 185 nm to 260 nm with a data pitch of 0.2 nm, a scan speed of 50 nm min^{-1}, a response time of 0.5 s and a bandwidth of 1 nm using quartz cells with a 1 mm path length, on a Jasco dichrograph equipped with a thermostatically controled cell-holder and connected to a computer for data acquisition. Each spectrum was recorded five times and averaged. Measurements were done at 25 °C and at 95 °C. The CD signal was corrected by buffer subtraction and converted to mean residue ellipticity. Data were acquired

from 10 µM samples in 50 mM sodium phosphate buffer, pH 7.0, in quartz cells with a 1 mm path length.

Thermal denaturation measurements

Thermal denaturation of αRep proteins was done in a JASCO spectropolarimeter equipped with a Peltier-type temperature controller with a heating rate of 1 K min^{-1}. Changes in $[\theta]_{222}$ of each protein were measured in the temperature range 35 – 95 °C (data pitch, 0.1 °C; response time, 1 s; bandwidth, 1 nm). All measurements were done with 10 µM αRep samples in 50 mM sodium phosphate buffer, pH 7.0. The data (*y*, *T*) were fit with a two-state denaturation model using the relation:

$$y(T) = y_N + sn(T) + \frac{\exp\left[-\left(\Delta H_m / R\right)\left(1/T - 1/T_m\right)\right]\left(A + \left(sd - sn\right)T\right)}{1 + \exp\left[-\left(\Delta H_m / R\right)\left(1/T - 1/T_m\right)\right]}$$

where y_N is the *y* intercept, *sn* is the initial slope, *sd* is the slope at the end of the denaturation, *A* is the amplitude of the denaturation transition, T_m (°C) is the midpoint of the thermal denaturation and *R* is the universal gas constant (kcal mol^{-1}).

Differential scanning calorimetry (DSC)

Thermal stability was studied by differential scanning calorimetry (DSC) with a MicroCal VP-DSC instrument with αRep proteins (0.23 – 1.25 mg mL^{-1}) in 50 mM sodium phosphate buffer, pH 7. Each measurement was preceded by a baseline scan with the standard buffer. Scans were done at 1K min^{-1} between 20 °C and 110 °C. The heat capacity of the buffer was subtracted from that of the protein sample before analysis. These corrected data were analyzed using a cubic spline as a baseline in the transition. Thermodynamic parameters ΔH_{cal} and ΔH_{vH} were determined by fitting the following equation to the data:

$$\Delta C_p(T) = \frac{K_d(T)\Delta H_{cal}\Delta H_{vH}}{\left[1 + K_d(T)\right]^2 RT^2}$$

where K_d is the equilibrium constant for a two-state process, ΔHvH is the enthalpy calculated on the basis of a two-state process and ΔHcal is the measured enthalpy.

Analytical size-exclusion chromatography

Analytical SEC was done with an ÄKTA Purifier (GE Healthcare) system using a Superdex™ 75 10/300 GL column (flow-rate 0.8 mL min^{-1}) equilibrated in 50 mM sodium phosphate buffer pH 7.0. For all the proteins analyzed, 100 µL of 3 mg mL^{-1} protein were injected onto the column. A solution containing standard globular proteins was injected as a control. As molar extinction coefficients were different for each protein, data were normalized relative to the maximum absorbance at 280 nm of each elution profile.

Crystallization and resolution of the structure

Both native and SeMet-labeled αRep-n4-a proteins were purified in 20 mM Hepes pH 7, 150 mM NaCl. A first crystal form (PEG-bound form) was obtained at 19 °C from a 1:1 (v/v) mixture of a 15 mg mL^{-1} protein solution with crystallization solution (18.5% PEG 1000, 6.5% glycerol, 100 mM Tricine, pH 8, 230 mM NaCl)**.** For data collection, the crystals were cryoprotected by transfer into the crystallization solution with progressively higher concentrations of glycerol up to 30% (v/v) and then flash-cooled in liquid nitrogen. The native and SeMet diffraction data for this crystal form were recorded on beamlines ID14-EH2 (ESRF, Grenoble, France) and Proxima-1 (synchrotron SOLEIL, France), respectively. A second crystal form (apo form) was obtained at 19 °C from a 1:1 (v/v) mixture of 15 mg mL^{-1} protein with crystallization solution (4% (v/v) MPD, 0.1 M sodium citrate, pH 5.6, 0.1 M $MgCl_2$). For data collection, the crystals were cryoprotected by transfer into the crystallization solution with progressively higher concentrations of MPD up to 30% (v/v) and then flash-cooled in liquid nitrogen. The diffraction data for this crystal form were recorded on beamline Proxima-1 (synchrotron SOLEIL, France).

The structure was determined by the multiple-wavelength anomalous dispersion (MAD) method using the anomalous signal from the selenium element. To do so, the diffraction data were recorded at the wavelengths corresponding to the peak, edge and remote of the selenium fluorescence spectrum. Data were processed using the XDS package.[41] The space group was $P2_1$ with two molecules per asymmetric unit. All the expected Se sites (two/monomer) were found in the 50 – 3 Å resolution range with the program SHELXD.[42] Refinement of the Se atom positions, phasing and density modification were done with the program SHARP.[43] The quality of the experimental phases allowed automatic building of all the α helices using the secondary structure recognition option of the ARP/WARP software.[44]

This model was then refined against the 2.15 Å native dataset using PHENIX[45] and then rebuilt with COOT. The first 15 amino acids corresponding to the streptag sequence: AWSHPQFEKAAAPLR followed by a three-alanine spacer and by the first three residues from the N-cap are not visible in the density map. The final model contains residues 16 – 200 and 20 – 201 from monomers A and B, respectively. In addition, 383 water molecules, two glycerol molecules and a large fraction of a PEG 1000 molecule could be modeled into the electron density maps.

The apo form was solved by molecular replacement using the program MOLREP,[46] and further refined to 1.8 Å resolution using PHENIX. The final model contains residues 6 – 196 and 139 water molecules.

Small-angle X-ray scattering

SAXS experiments

All proteins were studied at different concentrations in the range 0. 5–2 $g.L^{-1}$. A wider range of concentrations was used for αRep-*n*3-a (0.3–10.1 gL^{-1}) and αRep-*n*6-a (0.15–4.6 gL^{-1}). Most of the SAXS data were collected at the Swing beamline at SOLEIL, Saclay.[47] The sample-to-detector distance was 1.84 m, covering the range of scattering vector (q) from 0.008 $Å^{-1}$ to 0.4 $Å^{-1}$$q=4\pi sin\theta/\lambda$where 2θ is the scattering angle and λ (1. 033 Å) is the wavelength of the X-rays. The detector used was a CCD camera from Aviex. Twenty successive frames of 4 s exposure each separated by a 2 s pause were recorded for each protein solution and buffer. During exposure, the solution was contained in a quartz capillary 1.5 mm in diameter under vacuum. The solution was circulated continuously across the beam using an automated injection system. No radiation damage was detected under these conditions. Some SAXS data were collected on a commercially available small-angle X-ray instrument (Nanostar, Bruker AXS) adapted to a Microstar rotating anode X-ray source (CuKα; wavelength 1.54 Å). The instrument has an integrated vacuum in order to reduce background. The sample-to-detector distance was 662 mm so that intensities were recorded in the q-range 0.012 $Å^{-1} < q < 0.4$ $Å^{-1}$. Samples were enclosed in 2 mm diameter quartz capillaries inserted directly into vacuum. In both cases, samples were kept at a constant 15 °C.

Data analysis

SAXS data were normalized to the intensity of the incident beam, averaged and background-subtracted using the program package PRIMUS.[48] Intensities were put on an absolute scale using water scattering. Scattering intensities extrapolated to the zero angle $I(0)$ were first determined by Guinier approximation[49] from the low q-region of the scattering profiles, also yielding the value of the radius of gyration (R_g): $\ln I(q) = \ln I(0) - R_g^2 q^2 / 3$

The molecular mass of the scattering objects and consequently the oligomerization state of the proteins were derived from the value of $I(0)$ on an absolute scale. The distance distribution function $p(r)$ was calculated for each protein using the program GNOM.[50] Alternative values of R_g and $I(0)$ can be derived from $p(r)$. In all cases, they were in agreement with the values obtained using Guinier's law.

Protein models were built starting from the only crystal structure, that of αRep-*n*4-a. The *n*1, *n*2 and *n*3 models were obtained by suppressing from the αRep-*n*4-a structure the last three, two and one repeats, respectively while preserving the last C-terminal 36 residues. In the case of αRep-*n*6-a, two repeats were added in the center of the αRep-*n*4-a structure while keeping constant the relative orientation of two consecutive repeats. Calculated scattering curves from these models were finally fitted to the experimental curve by adding missing residues (10 in the N-terminal position and five in the C-terminal position) using the programs BUNCH,[51] SABBAG[52] and CRYSOL.[53] The experimental pattern used for fitting was obtained by extrapolation to infinite dilution of a set of curves recorded at decreasing concentrations in order to correct for the effect of interactions between proteins.

Protein Data Bank accession numbers

The atomic coordinates have been deposited in the Brookhaven Protein Data Bank with accession numbers 3LTJ for the apo form and 3LTM for the PEG-bound form.

Acknowledgements

We thank J. Perez and G. David (SOLEIL synchrotron) for help with the SAXS data collection and Magali Aumont-Nicaise for her help in DSC data recording and analysis. We thank Karine Blondeau for her help with SeMet labeling, and Jérome Cicolari and Anthony Doizy for preparation of crystals.

References

1 A. Skerra, Alternative non-antibody scaffolds for molecular recognition. *Curr. Opin. Biotechnol.*, **18** (2007), pp. 295–304.

2 H.K. Binz, P. Amstutz and A. Pluckthun, Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.*, **23** (2005), pp. 1257–1268.

3 H.K. Binz and A. Pluckthun, Engineered proteins as specific binding reagents. *Curr. Opin. Biotechnol.*, **16** (2005), pp. 459–469.

4 P.A. Nygren, Alternative binding proteins: affibody binding proteins developed from a small three-helix bundle scaffold. *FEBS J.*, **275** (2008), pp. 2668–2676.

5 A. Koide and S. Koide, Monobodies: antibody mimics based on the scaffold of the fibronectin type III domain. *Methods Mol. Biol.*, **352** (2007), pp. 95–109.

6 S. Schlehuber and A. Skerra, Anticalins as an alternative to antibody technology. *Expert Opin. Biol. Ther.*, **5** (2005), pp. 1453–1462.

7 H. Ebersbach, E. Fiedler, T. Scheuermann, M. Fiedler, M.T. Stubbs and C. Reimann, *et al.* Affilin-novel binding molecules based on human gamma-B-crystallin, an all beta-sheet protein. *J. Mol. Biol.*, **372** (2007), pp. 172–185.

8 A. Drevelle, M. Graille, B. Heyd, I. Sorel, N. Ulryck and F. Pecorari, *et al.* Structures of in vitro evolved binding sites on neocarzinostatin scaffold reveal unanticipated evolutionary pathways. *J. Mol. Biol.*, **358** (2006), pp. 455–471.

9 B. Heyd, F. Pecorari, B. Collinet, E. Adjadj, M. Desmadril and P. Minard, In vitro evolution of the binding specificity of neocarzinostatin, an enediyne-binding chromoprotein. *Biochemistry*, **42** (2003), pp. 5674–5683.

10 A. Drevelle, A. Urvoas, M.B. Hamida-Rebai, G. Van Vooren and M. Nicaise, *et al.* Disulfide bond substitution by directed evolution in an engineered binding protein. *ChemBiochem*, **10** (2009), pp. 1349–1359.

11 P. Forrer, M.T. Stumpp, H.K. Binz and A. Pluckthun, A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett.*, **539** (2003), pp. 2–6.

12 T.Z. Grove, A.L. Cortajarena and L. Regan, Ligand binding by repeat proteins: natural and designed. *Curr. Opin. Struct. Biol.*, **18** (2008), pp. 507–515.

13 A.V. Kajava, Review: proteins with repeated sequence – structural prediction and modeling. *J. Struct. Biol.*, **134** (2001), pp. 132–144.

14 Z. Pancer and M.D. Cooper, The evolution of adaptive immunity. *Annu. Rev. Immunol.*, **24** (2006), pp. 497–518.

15 H.K. Binz, M.T. Stumpp, P. Forrer, P. Amstutz and A. Pluckthun, Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.*, **332** (2003), pp. 489–503.

16 L.K. Mosavi, D.L. Minor and Z.Y. Peng, Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl Acad. Sci. USA*, **99** (2002), pp. 16029–16034.

17 H.K. Binz, P. Amstutz, A. Kohl, M.T. Stumpp, C. Briand and P. Forrer, *et al.* High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.*, **22** (2004), pp. 575–582.
18 A. Schweizer, H. Roschitzki-Voser, P. Amstutz, C. Briand, M. Gulotti-Georgieva and E. Prenosil, *et al.* Inhibition of caspase-2 by a designed ankyrin repeat protein: specificity, structure, and inhibition mechanism. *Structure*, **15** (2007), pp. 625–636.

19 G. Sennhauser, P. Amstutz, C. Briand, O. Storchenegger and M.G. Grutter, Drug export pathway of multidrug exporter AcrB revealed by DARPin inhibitors. *PLoS Biol.*, **5** (2007), p. e7.

20 M.T. Stumpp, P. Forrer, H.K. Binz and A. Pluckthun, Designing repeat proteins: modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J. Mol. Biol*, **332** (2003), pp. 471–487.

21 E.R. Main, K. Stott, S.E. Jackson and L. Regan, Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proc. Natl Acad. Sci. USA*, **102** (2005), pp. 5721–5726.

22 A.L. Cortajarena, T.Y. Liu, M. Hochstrasser and L. Regan, Designed proteins to modulate cellular networks. *ACS Chem Biol.*, **5** (2010), pp. 545–552.

23 M.E. Jackrel, A.L. Cortajarena, T.Y. Liu and L. Regan, Screening libraries to identify proteins with desired binding activities using a split-GFP reassembly assay. *ACS Chem. Biol.*, **5** (2010), pp. 553–562.

24 A.L. Cortajarena, F. Yi and L. Regan, Designed TPR modules as novel anticancer agents. *ACS Chem. Biol.*, **3** (2008), pp. 161–166.

25 F. Parmeggiani, R. Pellarin, A.P. Larsen, G. Varadamsetty, M.T. Stumpp and O. Zerbe, *et al.* Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *J. Mol. Biol.*, **376** (2008), pp. 1282–1304.

26 M.A. Andrade, C. Petosa, S.I. O'Donoghue, C.W. Muller and P. Bork, Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.*, **309** (2001), pp. 1–18.

27 F. Kippert and D.L. Gerloff, Highly sensitive detection of individual HEAT and ARM repeats with HHpred and COACH. *PLoS One*, **4** (2009), p. e7148.

28 O. Julien, I. Gignac, A. Hutton, A. Yee, C.H. Arrowsmith and S.M. Gagne, MTH187 from Methanobacterium thermoautotrophicum has three HEAT-like repeats. *J. Biomol. NMR*, **35** (2006), pp. 149–154.

29 F.B. Dean, J.R. Nelson, T.L. Giesler and R.S. Lasken, Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.*, **11** (2001), pp. 1095–1099.

30 S.L. Dahlroth, P. Nordlund and T. Cornvik, Colony filtration blotting for screening soluble expression in *Escherichia coli*. *Nat. Protoc.*, **1** (2006), pp. 253–258.

31 T. Cornvik, S.L. Dahlroth, A. Magnusdottir, M.D. Herman, R. Knaust, M. Ekberg and P. Nordlund, Colony filtration blot: a new screening method for soluble protein expression in *Escherichia coli*. *Nat. Methods*, **2** (2005), pp. 507–509.

32 D. Steiner, P. Forrer, M.T. Stumpp and A. Pluckthun, Signal sequences directing cotranslational translocation expand the range of proteins amenable to phage display. *Nat. Biotechnol.*, **24** (2006), pp. 823–831.

33 A.V. Kajava, What curves alpha-solenoids? Evidence for an alpha-helical toroid structure of Rpn1 and Rpn2 proteins of the 26 S proteasome. *J. Biol. Chem.*, **277** (2002), pp. 49791–49798.

34 D.I. Svergun, M.V. Petoukhov and M.H. Koch, Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.*, **80** (2001), pp. 2946–2953.

35 M. Socolich, S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner and R. Ranganathan, Evolutionary information for specifying a protein fold. *Nature*, **437** (2005), pp. 512–518.

36 T.J. Magliery and L. Regan, Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *J. Mol. Biol.*, **343** (2004), pp. 731–745.

37 M. Burg, E.P. Ravey, M. Gonzales, E. Amburn, P.H. Faix, A. Baird and D. Larocca, Selection of internalizing ligand-display phage using rolling circle amplification for phage recovery. *DNA Cell Biol.*, **23** (2004), pp. 457–462.

38 D. Christ, K. Famm and G. Winter, Tapping diversity lost in transformations – in vitro amplification of ligation reactions. *Nucleic Acids Res.*, **34** (2006), p. e108.

39 G.E. Crooks, G. Hon, J.M. Chandonia and S.E. Brenner, WebLogo: a sequence logo generator. *Genome Res.*, **14** (2004), pp. 1188–1190.

40 D. Steiner, P. Forrer and A. Pluckthun, Efficient selection of DARPins with sub-nanomolar affinities using SRP phage display. *J. Mol. Biol.*, **382** (2008), pp. 1211–1227.

41 W. Kabsch, Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *Acta Crystallogr. D*, **50** (1993), pp. 760–763.

42 T.R. Schneider and G.M. Sheldrick, Substructure solution with SHELXD. *Acta Crystallogr. D*, **58** (2002), pp. 1772–1779.

43 G. Bricogne, C. Vonrhein, C. Flensburg, M. Schiltz and W. Paciorek, Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallogr. D*, **59** (2003), pp. 2023–2030.

44 R.J. Morris, P.H. Zwart, S. Cohen, F.J. Fernandez, M. Kakaris and O. Kirillova, *et al.* Breaking good resolutions with ARP/wARP. *J. Synchrotron Radiat.*, **11** (2004), pp. 56–59.

45 P.D. Adams, R.W. Grosse-Kunstleve, L.W. Hung, T.R. Ioerger, A.J. McCoy and N.W. Moriarty, *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D*, **58** (2002), pp. 1948–1954.

46 A. Vagin and A. Teplyakov, MOLREP: an automated program for molecular replacement. *J. Appl. Crystallogr.*, **30** (1997), pp. 1022–1025.

47 G. David and J. Pérez, Combined sampler robot and high-performance liquid chromatography: a fully automated system for biological small-angle X-ray scattering experiments a the Synchrotron SOLEIL SWING beamline. *J. Appl. Crystallogr*, **42** (2009), pp. 892–900.

48 P.V. Konarev, V.V. Volkov, A.V. Sokolova, M.H.J. Koch and D.I. Svergun, PRIMUS: a windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.*, **36** (2003), pp. 1277–1282.

49 A. Guinier and G. Fournet, Small Angle Scattering of X-Rays, Wiley, New York (1955).

50 D.I. Svergun, Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.*, **25** (1992), pp. 495–503.

51 M.V. Petoukhov and D.I. Svergun, Global rigid body modelling of macromolecular complexes against small-angle scattering data. *Biophys. J.*, (2005).

52 J. Maupetit, R. Gautier and P. Tuffery, SABBAC: online structural alphabet-based protein backbone reconstruction from alpha-carbon trace. *Nucleic Acids Res.*, **34** (2006), pp. W147–W151.

53 D.I. Svergun, C. Barberato and M.H.J. Koch, CRYSOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.*, **28** (1995), pp. 768–773.

Conclusion

L'ensemble des résultats obtenus ont permis de valider notre approche dans la conception de protéines artificielles basées sur la répétition d'un motif *HEAT-like*. En effet la définition de la séquence du consensus et des « *cap* » et la technologie de construction des bibliothèques permettent d'obtenir une bibliothèque diverse. Ces protéines s'expriment à un taux important, se replient avec la structure attendue et sont très stables avec des températures de demi-dénaturation comprises entre 71 et plus de 95°C. De plus la diversité des séquences dans la banque correspond à celle codée sans biais apparent. Toutefois, l'analyse de séquence a révélé que 68% seulement des motifs sont codants et que le pourcentage des protéines codantes diminue avec l'augmentation du nombre de motifs insérés dans celle-ci. Par une expéience d'expression et de filtration sur boîte, on a montré aussi qu'uniquement 33% des variants s'expriment du fait de la proportion d'erreurs existant dans les modules. A ce niveau, notre priorité est devenue d'éliminer la source d'erreur de séquences dans le but d'augmenter la fraction des clones codants, en particulier ceux qui comprennent un nombre important de motifs. Le second objectif était d'améliorer la distribution des acides aminés et la diversité des positions hypervariables.

En plus du fait que ces protéines s'expriment bien, sont solubles et très stables, elles procurent aussi une large surface moléculaire qui peut contenir une grande variété de chaînes latérales dans le but d'élargir la gamme d'interactions éventuelles. En effet, des techniques de sélection, par phage display, ont été développées au sien de notre laboratoire. Et des sélections ont même été réalisées dans le but de valider la démarche suivie pour identifier des interacteurs affins et spécifiques à des cibles préalablement choisies.

Notre objectif était d'explorer les raisons possibles qui altéraient la qualité de construction de la banque, et d'évaluer des solutions permettant d'obtenir une banque de meilleure qualité. Le but essentiel étant de parvenir à la construction de la banque de deuxième génération, au sein de laquelle nous pourrions sélectionner des interacteurs spécifiques à des protéines cibles modèles. Ce travail d'amélioration puis de construction d'une banque de « seconde génération », les premières sélections d'interacteurs spécifiques et enfin la caractérisation des interactions et des complexes obtenus font l'objet du troisième volet de ce manuscrit.

Chapitre II :

Création d'une banque d'αRep de deuxième génération, sélection d'interacteurs pour des cibles choisies et caractérisation des interactions.

Table des matières

D. Sélection, identification d'interacteurs pour des molécules cibles et caractérisation des complexes obtenus

Abréviations

Amp: Ampicilline

BSA: *Bovine Serum Albumin*

DARPin: Designing Ankyrine Repeat Proteins

DO: Densité Optique

FS: Fraction Soluble

Gly: Glycérol

Glu: Glucose

h: Heure

IPTG : L'isopropyl β-D-1-thiogalactopyranoside

Kana : Kanamycine

min : Minutes

NCS : Néocarzinostatine

pb : paire de bases

PEG : Polyéthylène Glycol

RBS : *Ribosome Binding Site*

RCA : *Rolling Circle Amplification*

TBS : *Tris Buffer Saline*

TBST : TBS Tween 0.1%

Tet : Tétracycline

T7t : Terminateur du phage T7

A. Optimisation et construction du vecteur d'expression

Introduction :

A ce niveau du projet, nous avions pu valider la conception du motif *αRep* ainsi que la stratégie de construction de banque de variants. Mais nous avons aussi constaté un certain nombre d'anomalies que nous passerons en revue dans cette partie de la thèse et qu'il paraissait important d'améliorer. Nous allons tout d'abord détaillé ces améliorations méthodologiques: en premier lieu la conception et les tests correspondant au plasmide d'expression et d'exposition sur phages qui sera utilisé par la suite pour la nouvelle banque.

Nous avons choisi de ne pas renvoyer dans un chapitre spécial tous les éléments des matériels et méthodes, mais de les laisser proches des explorations réalisées et des résultats qu'ils ont permis d'obtenir. Toutefois, pour faciliter la lecture des chapitres qui suivent, les parties traitant des procédures expérimentales sont présentés de façon repérable, avec une police de caractère différente puis sont résumées dans l'annexe.

1. Le vecteur d'expression/display : phDiEx

Le vecteur utilisé a été conçu au laboratoire pour permettre l'expression d'une même protéine, soit à un niveau modéré et en fusion avec la protéine pIII du phage lors des expériences de *phage display*, soit à un niveau élevé et sans fusion avec pIII lors de la production de la protéine seule. Le passage de l'un à l'autre mode d'expression ne nécessite pas d'étape de sous clonage. En effet, l'expression sur phage ne nécessite pas un promoteur puissant et elle est habituellement réalisée sous contrôle d'un promoteur lactose. La fusion *αRep*-pIII est due à la suppression du codon stop (tag) situé entre la protéine à exposer et la protéine pIII. Pour cela, les expériences de production de phages sont réalisées dans des souches (type *XL1 blue*) contenant un gène suppresseur de non-sens (supE). Pour l'utilisation du vecteur en mode expression de protéine isolée, le promoteur T7 est utilisé et permet une transcription beaucoup plus efficace dans des souches exprimant la T7 polymérase (type *BL21 DE3*). Ces souches ne permettent pas la suppression des codons tag et la traduction s'arrête donc au codon stop situé entre la protéine exposée et la protéine pIII.

Dans la version de *phDiEx* précédemment utilisée au laboratoire pour des projets de *phage display* utilisant une autre protéine la Néocarziniostatine (NCS), le vecteur comportait une séquence d'export *ompT* permettant la sécrétion de la protéine dans le périplasme, pour le *phage display* comme pour l'expression. Il comportait également un *Strep-tagII* situé entre la

séquence d'export et la protéine exposée, et une séquence hexahistidine (His_6) coté C-terminal de la protéine exprimée, juste avant le codon stop suppressible (Fig. 1).

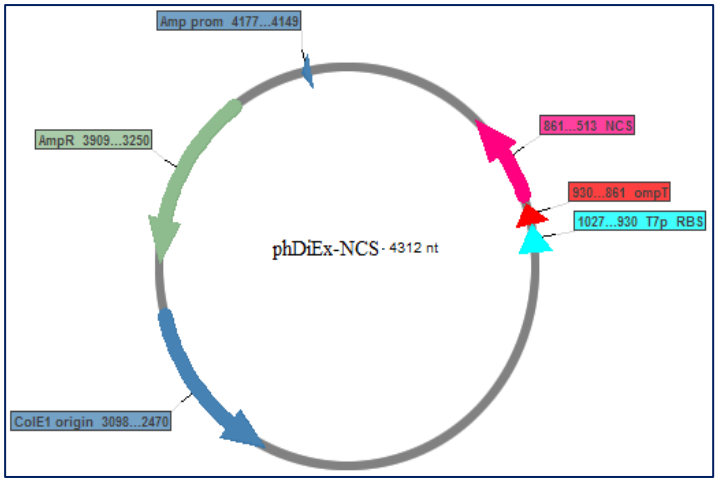


Fig. 1 : Plasmide *phDiEx* contenant le gène codant pour la NCS.

A partir des tests préliminaires d'expression en milieu liquide de clones d'une ankyrine (*Darpin)* synthétique choisie comme modèle de protéines très stables, nous avons constaté que l'expression à partir de ce vecteur (*phDiEx*) n'était pas aussi favorable avec cette protéine que ce qui avait été obtenu précédemment. En effet, lors d'un projet précédent réalisé avec une autre protéine, de stabilité moyenne, la Neocarzinostatine (ou NCS), ce vecteur permettait une exposition correcte de la protéine sur phages. Par ailleurs, un taux d'expression très élevé (>50mg/L) était obtenu par auto-induction et sécrétion de la protéine dans le milieu de culture. Des expériences ont alors été réalisées en prenant comme modèle de protéines thermostables une ankyrine dont la séquence est très proche d'une des premières ankyrines artificielles décrite par l'équipe de Plückthun : la *Darpin* E3-5. Dans les mêmes conditions d'expression (auto-induction) que ce qui était utilisé pour la NCS, la sécrétion de la *Darpin* semble peu efficace (typiquement < 10 mg/L). D'autres tests d'expression réalisés dans des souches d'expression (*BL21DE3* et dérivés) mais dans des conditions d'induction conventionnelles, par addition d'IPTG, montrent, en mode expression, un taux d'expression faible et une expression basale de la *Darpin* même en absence l'IPTG. Enfin, en mode *display*, le taux d'exposition de la *Darpin* sur phages paraissait très faible, et nettement inférieur à ce qui était obtenu au préalable avec la NCS.

Des tests ont été réalisés dans le but d'optimiser la séquence d'export pour parvenir à une meilleure expression et exposition sur phages des protéines stables, conditions qui pourraient être ensuite applicables aux *αRep*. Dans ce cadre, l'équipe a changé la séquence d'export *ompT* par une autre séquence d'export : *DsbA* (Fig. 2) de type SRP qui venait d'être

décrite pour permettre une exportation plus efficace de protéines stables. Les tests montraient effectivement que le niveau d'exposition de la *Darpin* modèle était très nettement amélioré avec cette séquence d'exportation qui a donc été retenue pour la suite.

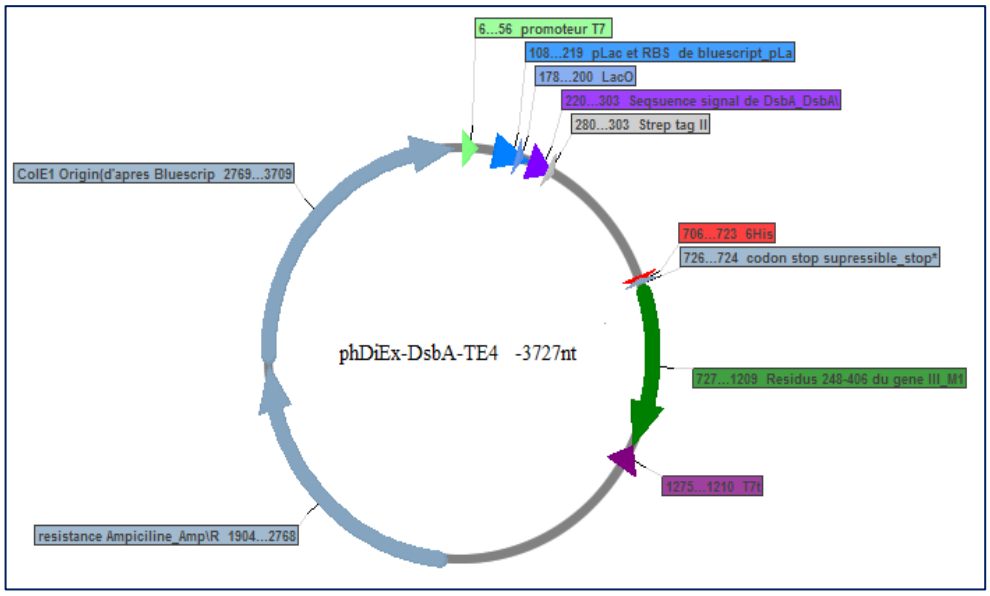


Fig. 2 : Plasmide *phDiEx* contenant les gènes codants pour le promoteur T7, LacO, la séquence d'export *DsbA*, deux étiquettes Sterp-tagII (coté N-terminal de la protéine à exprimer) et His_6 (coté C-terminal), le codon Stop suppressible et une partie du gène III du phage M13.

Concernant l'expression « hors phage » des *αRep* avec la version préalablement développée du vecteur *phDiEx*, les résultats montraient une expression plutôt faible et semblaient moins favorables que ce qui était obtenu avec la protéine précédente (NCS). Pourtant lorsque les séquences codantes d'*αRep* sont sous-clonées dans un vecteur d'expression cytoplasmique de type pQE31 (Fig. 3), l'expression est très efficace et toutes les *aRep* testées s'expriment sous forme soluble en concentration élevée.

La faible efficacité d'expression dans *phDiEx* n'est donc pas un problème relié à l'agrégation des *aReps* ou à leur sensibilité éventuelle aux protéases, mais plutôt un problème de transcription, de stabilité du messager ou de traduction.

La construction d'une nouvelle bibliothèque suppose alors la fixation, dès le départ, du choix d'un vecteur dans lequel cette bibliothèque sera construite. Il était important d'examiner et si possible d'améliorer ses caractéristiques avant d'entreprendre la construction d'une nouvelle bibliothèque.

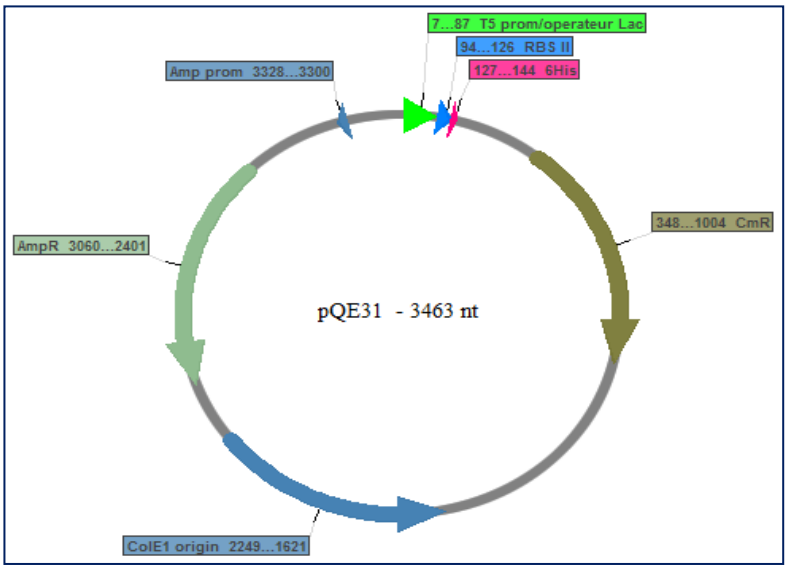


Fig. 3 : Carte du plasmide d'expression pQE31.

Nous nous sommes intéressés alors aux causes éventuelles de ce faible taux d'expression de *phDiEx* dans le but d'y remédier. Nous avons tout d'abord vérifié si une anomalie dans la séquence du promoteur (LacO ou T7) ou du terminateur non détectée aurait pu survenir au fil des constructions/amplifications du vecteur. Cela aurait pu expliquer une variabilité observée sur des différences d'expression lors d'expériences réalisées avec un écart de temps important. Les séquences des régions 5' (promoteur T7, RBS) et 3' (terminateur T7) du vecteur ont été vérifiées et ne présentent pas d'anomalies.

Nous avons alors souhaité reprendre des tests d'expression comparés dans le but d'évaluer le rôle éventuel de différents paramètres sur l'efficacité d'expression.

- Nature de la protéine et/ou de la séquence d'exportation périplasmique.
- Conditions d'induction
- Influence de la région codant la protéine pIII
- Influence de la distance entre promoteur et RBS.

2. Le taux d'expression faible est-il dû à l'existence de la protéine pIII du phage M13 ?

L'existence de la protéine pIII du phage M13 en fusion avec la protéine (*αRep*) exprimée donne un ARN messager long, ce qui peut favoriser leur dégradation par les RNAases, surtout lorsqu'une partie significative de ce messager n'est pas traeduite comme c'est le cas en mode expression dans les souches non suppressives. Pour évaluer cet éventuel effet, nous avons

comparé l'expression d'une même *αRep* de référence (αRep-n4-a) dans le vecteur *phDiEx*, dans le vecteur *phDiEx* modifié pour ne plus comporter la protéine pIII du phage M13 et dans un vecteur de sécrétion disponible dans notre laboratoire (*pSec*). Dérivé d'un *pET12a*, c'est à dire reposant sur un promoteur T7, ce plasmide *pSec* a été beaucoup utilisé au laboratoire pour l'expression d'une autre famille de protéines dérivées de la Neocarzinostatine (NCS). De même, la surexpression de ces NCS par le vecteur de display/expression a été caractérisée au laboratoire, notamment lors de la thèse d'Antoine Drevelle[3]. Les résultats d'expression de la NCS dans *phDiEx* pouvaient nous servir de repère d'efficacité et nous guider dans les modifications à apporter au plasmide pour la construction de notre nouvelle banque.

2.1. Expression et Purification comparative de la NCS dans *phDiEx* et *pSec*

Les deux plasmides pSec et phDiEx contenant le gène la protéine NCS ont été utilisés pour transformer des bactéries BL21DE3. Deux colonies des bactéries transformées ont été utilisées pour inoculer 50 mL 2YT+Amp+Glu puis incubées o.n. à 37°C. Ces précultures sont par la suite utilisées pour ensemencer 500 mL 2YT+Amp. Les cultures sont incubées 72 h à 30°C dans le but d'exprimer la NCS par auto-induction puis sécrétion dans le milieu de culture. Les cellules bactériennes ont été récupérées par centrifugation (20 min, 4000 g) et le surnageant est retenu pour la purification des protéines. Les protéines des deux surnageants ont été précipitées au sulfate d'ammonium. Les culots protéiques sont repris dans l'eau et par la suite dialysés. Les protéines solubilisées sont par la suite purifiées par affinité sur une colonne de Nickel et éluées dans un volume identique.
Une analyse des différentes fractions de la purification a été réalisée par SDS page sur gel d'acrylamide 15% (Fig. 4).

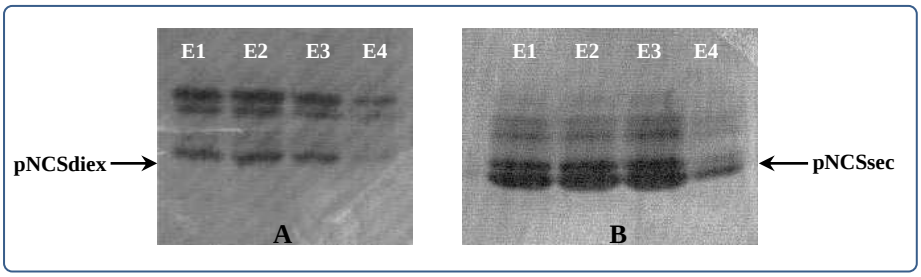


Fig. 4 : Analyse *SDS-page* des fractions (E_1, E_2, E_3 et E_4) issues de la purification sur Colonne Ni-NTA des surnageants de culture de bactéries transformées avec les 2 plasmides d'expression : *phDiEx* (A) et le *pSec* (B).

[3]Thèse de doctorat d'Antoine Drevelle : Evolution Dirigée de la néocarzinostatine : ingénierie d'une ossature protéique alternative aux anticorps. Université Paris-Sud11 année 2006-2007.

A partir de ces gels d'acrylamide, on retrouve avec la NCS les résultats attendus : l'expression de la NCS par auto-induction est très efficace, la protéine est dans ces conditions secrétée dans le milieu extérieur, vue sur gel B. La présence de la séquence codant la protéine pIII, non traduite, semble diminuer partiellement l'expression de la NCS, mais l'expression reste efficace. Dans les mêmes conditions l'*αRep* n'est pas efficacement exprimée.

Nous avons alors comparé l'expression d'une *αRep* dans ce plasmide de référence (ossature type pNCSsec) avec celle du vecteur de display/expression *phDiEx* en présence et en absence de la protéine de M13. La réalisation de ce test d'expression en milieu liquide a nécessité :

- le clonage du gène de l'αRep-n4-a dans le plasmide *pSec*.
- l'élimination du gène codant pour la pIII dans le plasmide *phDiEx*.
- l'expression des différents vecteurs dans une souche d'*E.coli* (*BL21DE3*).
- l'analyse par *Western Blot*.

2.2. Clonage du gène de αRep-n4-a dans le vecteur *pSec*

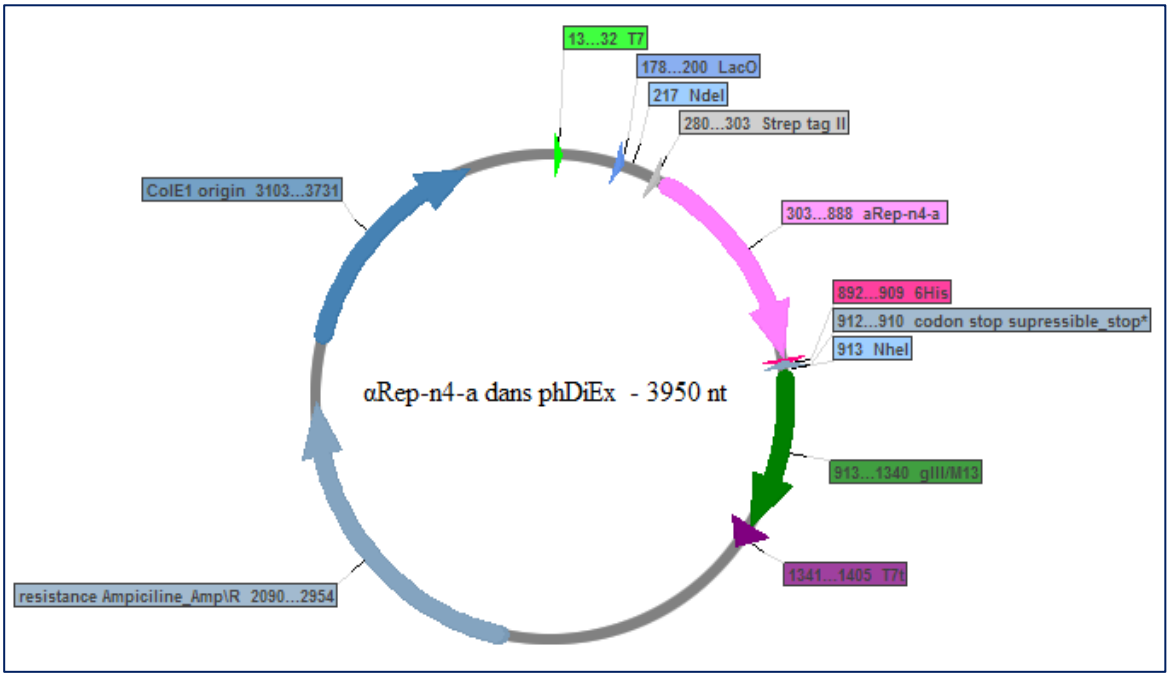


Fig. 5 : Le plasmide *phDiEx* avec αRep-n4-a avec les sites de restriction utilisés pour le clonage : NdeI-NheI.

Le clonage de l'αRep-n4-a (Fig. 5) s'est fait par le biais des enzymes de restriction NheI et NdeI. En premier lieu, les vecteurs pSec et phDiEx ont été digérés NheI-NdeI pendant 1 h à 37°C. Les produits de la double digestion ont été, par la suite, déposés sur un gel d'agarose 2 % pour le phDiEx-αRep et un gel d'agarose 1 % pour le pNCSsec (Fig. 6). Les bandes correspondant à la région codant la protéine (αRep) pour le phDiEx-αRep (≈ 600pb) et celle correspondant au vecteur pour le pNCSsec ont été découpées et l'ADN purifié du gel par le kit ExtractII de chez Macherey-Nagel. Les deux fragments ont été ligués o.n. à 16°C. Le produit de la ligation a été finalement purifié puis utilisé pour transformer des bactéries XL1Blue MRF' électrocompétentes. Des clones ont été pris au hasard pour vérifier ceux qui ont bien intégré l'insert. Par la suite, nous avons vérifié la séquence de αRep-n4-a par séquençage.

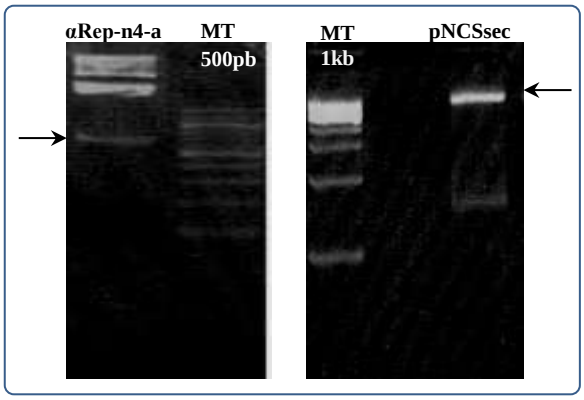


Fig. 6 : Profil de la digestion NdeI-NheI des 2 plasmides *phDiEx* et *pSec*: (A) Profil de digestion de *phDiEx* : avec la bande de 600pb correspondant à l'αRep-n4-a. (B) Profil de digestion *pSec* : la bande de 3.7kb correspondant au plasmide.

2.3. Construction du plasmide pαRep-n4-adiex sans le gène de la protéine pIII du phage M13

Au départ, nous avons tenté d'éliminer la séquence du gène III par simple digestion de site compatible de NheI et StyI (Fig.7).

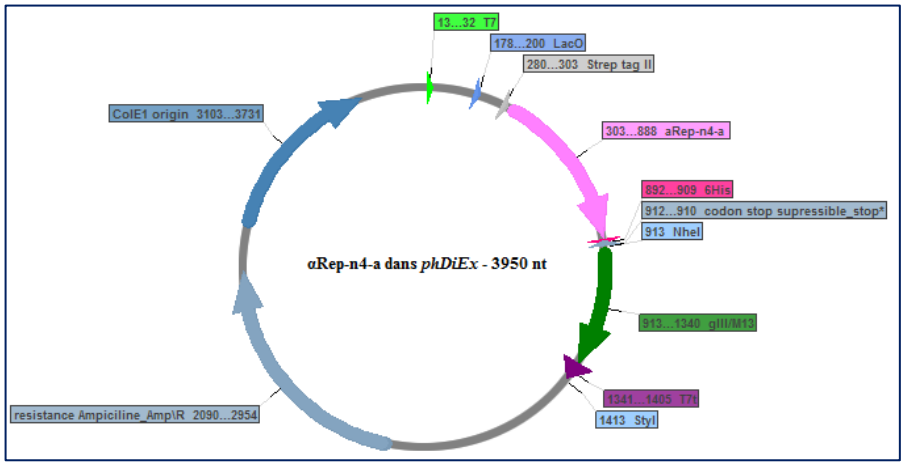


Fig. 7: Plasmide *phDiEx* avec le gène codant pour l'αRep-n4-a, le gène III et les sites de restrictions NheI et StyI.

En analysant la séquence et l'emplacement des sites sur le plasmide, nous avons remarqué que le site de StyI est localisé au niveau de la séquence du terminateur T7 (T7t). Ainsi, nous avons opté pour l'introduction d'un site NheI entre le gène gIII et le T7t par mutagénèse dirigée.

Nous avons alors utilisé les oligonucléotides suivants :

**site NheI-for :5'*GTAATAAGGAGTCTTAAGCTAGCTAAGATCCCTAGGATAACCCC*3'*

**site NheI-rev : 5'*GGGTTATGCTATTTAGCTAGCTTAAGACTCCTTATTAC*3'*

Après la PCR optimisée du plasmide phDiEx avec ces 2 primers, le produit de l'amplification a été utilisé pour l'électroporation de cellules XL1Blue MRF'. Des clones isolés ont été choisis pour extraire le plasmide et vérifier l'insertion du site NheI (Fig. 8).

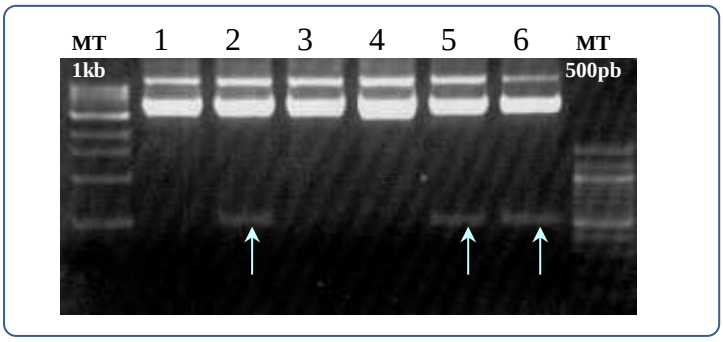


Fig. 8 : Profil de la digestion NheI des 6 clones pris au hasard après PCR : les clones 1, 3 et 4 montrent le profil d'un plasmide linéarisé (ayant un seul site NheI) alors que les clones 2, 5 et 6 montrent le profil d'un plasmide doublement digéré. Ils ont alors intégré le deuxième site NheI et la digestion libère le fragment d'ADN de 500pb correspondant au gène gIII.

Une quantité de plasmide plus importante des clones 2, 5 et 6 a été digérée avec NheI puis déposée sur un gel d'agarose 1%. Parmi les 2 fragments d'ADN (500 pb et 3.5 kb) séparés sur gel, la bande du plasmide de 3.5 kb a été purifiée puis l'ADN a été recircularisé par simple ligation. Les plasmides ont été, par la suite, utilisés pour électroporer des bactéries XL1Blue MRF'. Plusieurs clones ont été utilisés pour extraire les plasmides et vérifier l'absence du gène gIII par digestion NheI-BglII. 7 clones ont été retenus et par séquençage nous avons vérifié la séquence de l'αRep-n4-a sans gIII : plasmide phDiEx-ΔgIII.

2.4. Test d'expression de la αRep-n4-a en milieu liquide comparant *phDiEx*, *phDiEx-ΔgIII* et *pSec*

Le test d'expression en milieu liquide a été réalisé en deux étapes :

- Expression des protéines
- Analyse par *Western Blot*

a. Expression des protéines

Les plasmides correspondant à chaque construction ont été utilisés pour transformer des bactéries de la souche BL21DE3. Des précultures ont été réalisées par incubation d'une colonie dans 10 mL 2YT+Amp+Glu o.n. à 37°C et 220 rpm. Ces précultures ont été utilisées pour inoculer des cultures de 20 mL 2YT+Amp à DOi=0.1 qui sont par la suite incubées à 37°C et 220rpm. A une DO_{600}*= 0.6, l'expression a été induite par l'ajout d'IPTG. Après 4h d'induction les bactéries sont récupérées par centrifugation (10 min, 4000 g). Elles sont, par la suite, lysées dans le tampon B-PERII (Roche). Des échantillons du lysat total et de la fraction soluble ont été dénaturés par ajout de tampon SDS et chauffage 3 min à 100°C puis utilisés pour l'analyse par Western Blot.*

b. Analyse par Western Blot

Les échantillons dénaturés ont été déposés sur gel acrylamide 15% puis séparés par électrophorèse. Les protéines ont été transférées sur une membrane de nitrocellulose qui a été révélée. La membrane est tout d'abord bloquée dans du TBST BSA 3% o.n. sous agitation par la suite elle est incubée dans 10 mL d'une solution de TBST contenant un anticorps anti-His-tag (dilution 1/10000). Une fois la membrane lavée, elle est incubée dans une solution de TBST contenant l'anticorps secondaire : un anticorps anti-souris fluorescent à la dilution 1/50000. La membrane traitée est finalement scannée dans le scanner Odyssée à une longueur d'onde d'excitation de 700 nm (Fig. 9).

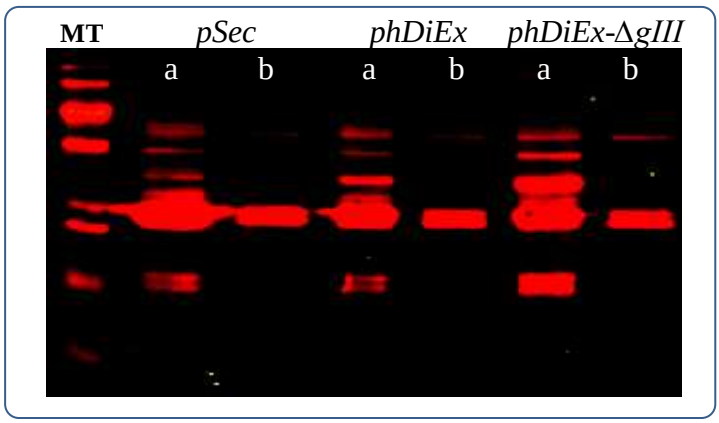


Fig. 9 : Test d'expression en milieu liquide (*Western blot*) comparant l'expression de l'αRep-n4-a avec les 3 constructions *pSec*, *phDiEx* et *phDiEx-ΔgIII* où les puits a et b correspondent respectivement à l'extrait total et la fraction soluble des culots bactériens lysés.

Nous avons analysé l'expression de l'*αRep*, dans les différentes constructions, par *Western blot*. Par comparaison au test d'expression préalable de la NCS, l'*αRep* est plus faiblement exprimée que la NCS. Plutôt qu'utiliser une coloration directe comme précédemment, nous avons donc comparé l'expression après révélation par *Western blot* mieux adapté que la coloration directe à l'évaluation des taux d'expression faibles.

La comparaison des fractions totales et solubles des différentes constructions montre que la protéine (*αRep*) est exprimée majoritairement sous forme soluble. La comparaison du taux d'expression, à partir des fractions solubles, des différentes constructions ne montre pas de différence significative en particulier en présence ou en absence de la pIII du phage M13.

En conclusion le taux relativement faible d'expression des *αRep* n'est pas augmenté si on supprime la partie non traduite (dans des souches type *BL21 DE3*) en 3' du messager de la pIII.

3. Optimisation du promoteur du vecteur et introduction d'un outil de criblage de la banque : pLac-T7prom-Flag-tag

Lors des expériences précédentes, nous avons vérifié que le faible taux d'expression des *αRep* ne provient pas de la présence de la protéine pIII du phage M13 et de la séquence non traduite qu'elle entraine. Nous nous sommes alors intéressés à l'organisation de la séquence non traduite en 5' et avons comparé l'efficacité en mode expression de deux plasmides qui diffèrent par l'inversion de l'ordre des deux promoteurs dans leur séquence. Dans la version préalablement utilisée du plasmide, le promoteur T7 se trouve positionné 200 nucléotides en amont de l'origine de traduction. Le promoteur lactose se trouve dans une configuration classique plus proche de la séquence codante et conduit donc à un messager comportant une séquence 5' non traduite relativement courte. Les séquences non traduites longues peuvent favoriser la dégradation du messager mais également peuvent conduire à l'établissement de structures secondaires très préjudiciables à l'efficacité de la traduction lorsqu'elles recouvrent le site d'initiation de la traduction. L'idée était donc de réduire les risques de dégradation ou de structuration du messager produit sous contrôle du promoteur T7 en plaçant celui-ci plus proche du site de fixation des ribosomes, comme c'est en général le cas dans les constructions basées sur ce promoteur. Cela conduit à éloigner le promoteur lactose utilisé en mode *display*, ce qui pourrait être défavorable à l'expression en mode *display*. Cependant le niveau d'expression requis en mode *display* est nettement moins élevé que celui requis en mode expression et il paraissait donc préférable d'optimiser prioritairement le placement du promoteur T7 plutôt que le promoteur Lac.

Nous avons conçu puis construit un vecteur appelé pLacT7, similaire à *phDiEx*, qui comporte une région promotrice réarrangée.

Nous avons, tout d'abord, réalisé ces deux types de construction avec une *αRep* de référence et par la suite vérifié pour chaque construction l'efficacité de l'expression et l'efficacité de l'exposition sur phages (taux de *display*).

3.1. Inversion des promoteurs et introduction du *Flag-tag* dans des clones références

Dans le but de construire le nouveau plasmide pLacT7, présenté dans paragraphe précédent, nous avons conçu puis synthétisé un plasmide contenant la séquence du promoteur pLac suivie par celle du promoteur T7 et celle du Flag-tag (Fig. 10) qui sera par la suite incorporée, par sous-clonage, dans le *phDiEx* à la place de la séquence codant le T7 pLac *Strep-tagII*.

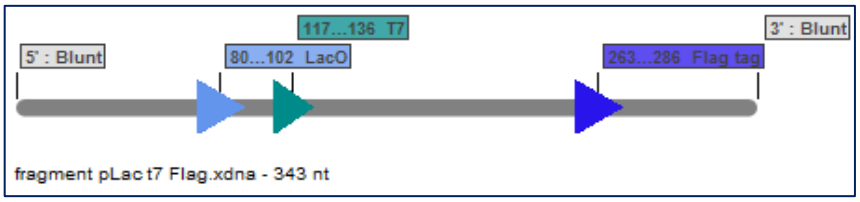


Fig. 10 : La carte du gène synthétique utilisé pour inverser les promoteurs et introduire le *Flag-tag*.

Nous avons changé le promoteur de deux *αRep* dont on connaît les propriétés d'expression : la αRep-n4-a et la αRep-n1-a, dans le but d'effectuer des tests d'expression et des tests d'exposition sur phages. Ceci est réalisé par clonage *via* les enzymes BglII-NotI (Fig. 11) comme décrit dans (I-2).

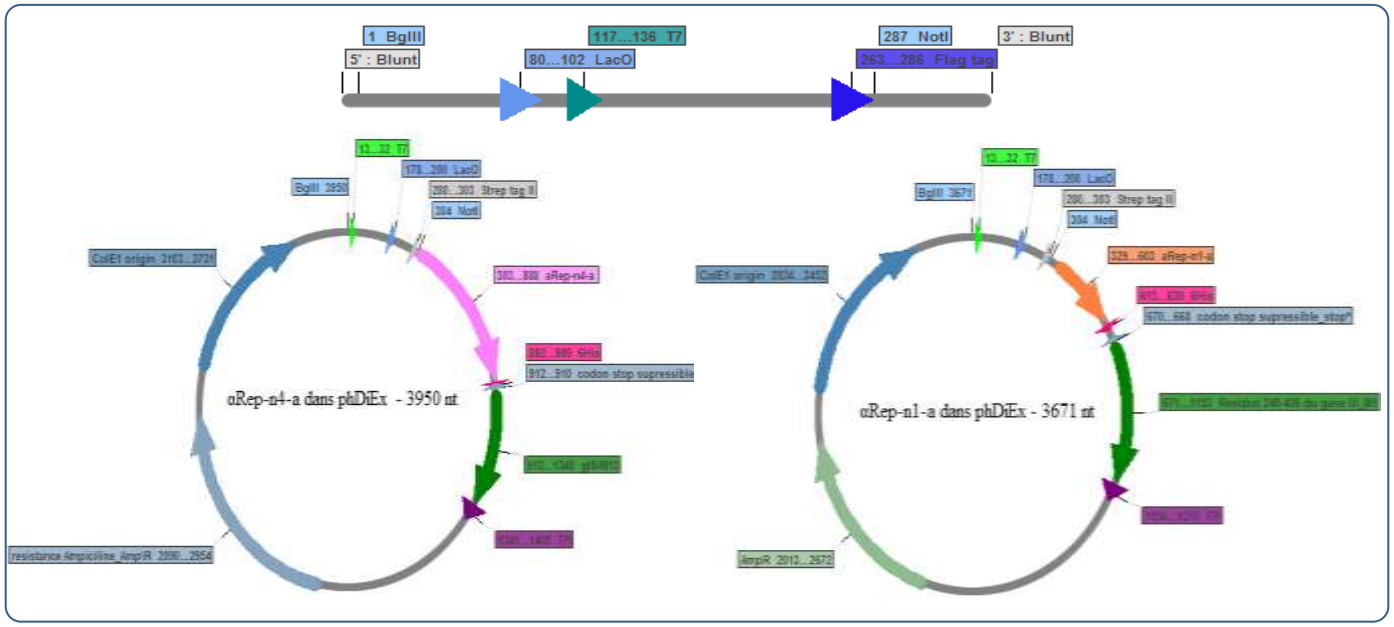


Fig. 11 : Carte des plasmides *phDiEx* de αRep-n4-a, αRep-n1-a et le plasmide synthétique montrant les sites de restriction qui ont servi pour le clonage.

En plus de l'inversion des promoteurs et l'introduction du *Flag-tag*, ce gène synthétique nous procure l'avantage de passer d'une expression périplasmique à l'expression cytoplasmique par simple digestion EagI qui élimine la séquence d'export. Des clones issus du clonage ont été séquencés puis alignés.

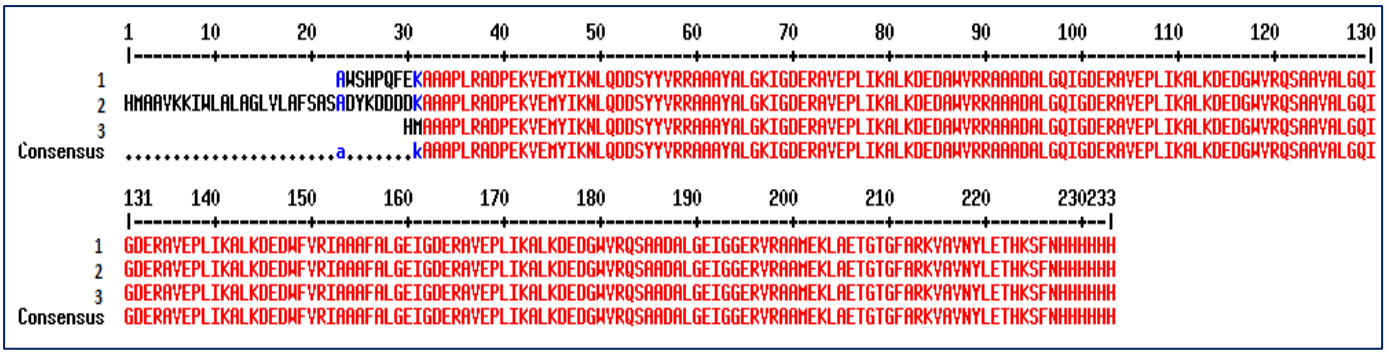


Fig. 12 : Alignement des séquences du clone αRep-n4-a obtenu après le clonage (2) [avec la séquence d'export et le nouveau promoteur] et du clone αRep-n4-a après digestion EagI (3) [sans la séquence d'export mais avec le nouveau promoteur].

L'alignement de séquence (Fig. 12) montre que les clones obtenus conservent la séquence de l'αRep-n4-a (1). Le clone (2) a la séquence de l'αRep-n4-a avec le *Flag-tag* (DYKDDDDK) au lieu du *Strep-tagII* (WSHPQFEK). Il a alors intégré le gène pLac-T7prom-*Flag-tag*. Le clone 3 a la séquence de αRep-n4-a avec le Flag-tag sans la séquence d'export : cette construction sera utilisée pour tester l'expression cytoplasmique (*cytex*). Ces 3 clones ont été utilisés pour le test d'expression en milieu liquide.

3.2. Test d'expression en milieu liquide

Ce test vise la comparaison de l'expression de l'αRep-n4-a dans 4 constructions différentes : ***pQE***, ***phDiEx***, ***pLac-T7prom-Flag-tag***, ***pLac-T7prom-Flag-tag-cytex***. Il a été réalisé comme décrit dans (I-4).

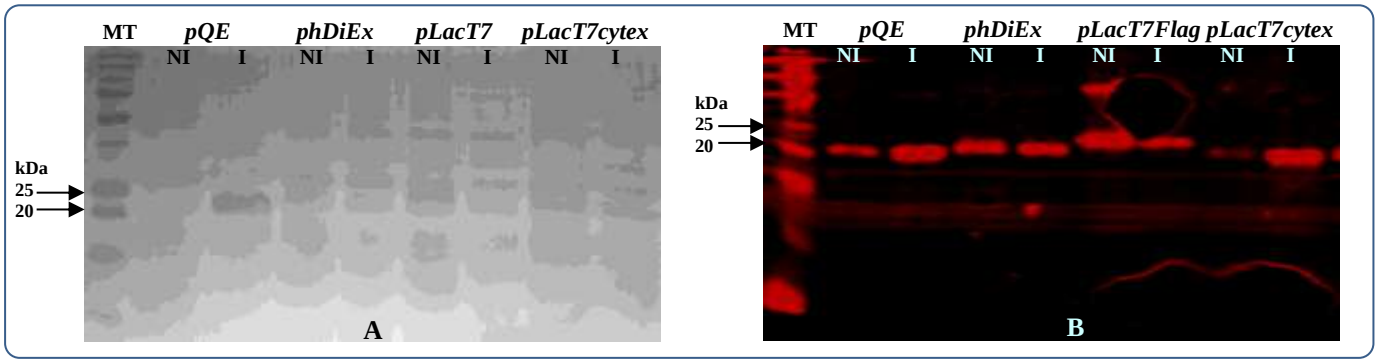


Fig. 13 : Test d'expression en milieu liquide comparant les différentes constructions plasmiques. A : Analyse *SDS page* des factions solubles de lysats bactériens des différentes cultures non-induites (NI) et induites (I). B : La membrane de nitrocellulose révèle spécifiquement l'αRep-n4-a, par le biais de son *His-tag*, à partir des fractions solubles des lysats bactériens des différentes cultures non-induites (NI) et induites (I).

Ce test d'expression en milieu liquide, montre, d'une part, que le meilleur taux d'expression est celui obtenu avec le plasmide d'expression *pQE*. Ce résultat apparait clairement lorsque le gel est révélé au bleu de Coomassie, cette méthode de coloration permettant d'évaluer la présence d'une protéine très fortement exprimée. L'identification de la protéine recherchée sur ce type de révélation étant plus ambiguë lorsque le niveau d'expression est plus faible, nous avons également révélés ces échantillons par un *Western blo*t. La très forte expression obtenue avec *pQE* y apparait moins clairement, du fait d'une saturation du signal obtenu avec une quantité de protéine inférieure à la quantité saturant la coloration au bleu de Coomassie. Nous remarquons sur le *Western blot* que le taux d'expression est similaire entre *phDiEx* et pLacT7prom-*Flag-tag*. Contrairement à ce que nous espérions en réalisant cette construction, le repositionnement du promoteur T7 et la diminution de la partie 5' non traduite du messager n'influe pas, de façon décisive, sur le niveau d'expression obtenu et ne permet pas d'atteindre le niveau d'expression obtenu avec *pQE*. Le processus d'exportation vers le périplasme peut aussi constituer une étape limitant l'efficacité d'expression. Les résultats obtenus avec la construction délétée de la séquence d'export (*cytex*) ne conduisent pas à un niveau d'expression aussi élevé que ce qui est obtenu avec le système d'expression *pQE*.

Par ailleurs, une expression avant induction (ou « fuite ») est observée avec toutes les constructions mais semble nettement plus faible avec *pQE* ainsi qu'avec le plasmide conçu pour une expression cytoplasmique. Pour résumer la construction comportant une inversion des promoteurs T7 lac ne conduit pas à une expression périplasmique plus efficace. Le taux d'expression est un peu plus élevé en expression cytoplasmique que périplasmique et même dans ces conditions il reste plus faible que celui obtenu avec *pQE*. Ces résultats ont été aussi confirmés pour le clone αRep-n1-a.

En conclusion, les améliorations recherchées dans l'efficacité d'expression n'ont pas été réellement obtenues. Le niveau d'expression obtenu avec un autre système (*pQE*) n'a pu être atteint ni par la réorganisation de la région promotrice, ni par la délétion d'une séquence non traduite en 3', ni par la suppression de la séquence d'export. Nous avons donc été conduits à nous accommoder de ces limites ce qui impliquera ultérieurement d'utiliser une étape de sous clonage dans un système type *pQE* lorsqu'une très forte expression sera nécessaire.

Nous avons tout de même décidé d'utiliser le pLacT7prom comme vecteur pour la construction de la banque d'*αRep*. Ce plasmide, bien qu'il ne présente pas de propriétés d'expression améliorées, a l'avantage de comporter un *Flag-tag* qui sera un moyen de sélectionner, avec une affinité plus élevée que ne le permettait le *Strep-tagII*, les phages exprimant une protéine qui comporte cette étiquette. Enfin cette construction permet d'exciser très simplement la séquence d'exportation périplasmique grâce aux deux sites de restriction EagI situés de part et d'autre de cette séquence, ce qui permet de passer plus simplement d'une expression périplasmique à une expression cytoplasmique. Il reste cependant un paramètre à vérifier : le taux de *display*. Ce dernier est un paramètre important pour les sélections et le criblage de la banque par exposition sur phages.

3.3. Comparatif du taux de display entre le vecteur *phDiEx* et le vecteur *pLacT7prom*

Au sein de notre équipe, des essais de purification des *αRep* par le *Strep-tagII* ont été réalisés et semblaient peu efficace et irréguliers. Il est possible que l'affinité modérée du *Strep-tagII* pour la streptactine (K_D de l'ordre du µM) soit en cause, en particulier avec des protéines monomériques ne manifestant pas d'effets d'avidité. De plus, des analyses par spectrométrie de masse ont alors montré que les protéines sont exprimées avec des séquences du *Strep-tagII* partiellement tronquées. Nous avons alors décidé de substituer cette étiquette pour introduire le Flag-tag. Celui-ci est supposé avoir une affinité plus élevée pour l'anticorps *anti-Flag* que le *Strep-tagII* pour la streptactine, ce qui permettrait de disposer d'un meilleur outil de criblage de la banque. Il est en effet important de développer une méthode permettant de « filtrer » les séquences sans erreur, c'est-à-dire de retenir sélectivement la sous-population de clones qui expriment une protéine codante en *phage display* et d'éliminer ceux qui comportent des modules erronés. Cela permettra ultérieurement de recréer de la diversité en recombinant les modules codants entre eux.

Mais avant cela, il fallait vérifier l'efficacité d'exposition des protéines à la surface des phages et pour cela mesurer un « taux de *display* ». Ceci est réalisé par le comptage des phages exposant l'*αRep* parmi les phages totaux produits dans la culture.

Cette mesure est réalisée en utilisant une population homogène de phages exposant une αRep (l'αRep-n4-a) dont la séquence est codante. La protéine, si elle est exposée sur

phage, doit comporter à son extrémité N-terminale, accessible sur le virion le peptide *Flag*. Ces particules virales comportant un *Flag-tag* peuvent être capturées sur une plaque ELISA recouverte d'anticorps *anti-Flag*, et les phages retenus spécifiquement sont élués à pH acide. On peut quantifier les phages initialement ajoutés (*Input*) ou les phages retenus puis élués (*Output*) par infection de bactéries sensibles et numération de colonies

Des bactéries XL1Blue MRF' sont tout d'abord transformées, indépendamment, avec le vecteur phDiEx et le vecteur pLacT7prom codant pour αRep-n4-a. Une colonie issue de chaque transformation est prélevée puis incubée o.n. à 37°C et 220 rpm dans 10 mL 2YT+Amp+Tet+Glu. A partir des précultures, des cultures de 20mL 2YT+Amp+Tet ont été préparées à DOi=0.1. A DO=0.6, les bactéries sont infectées par le phage helper M13KO7 et incubées à 37°C pendant 30' sans agitation puis 30 min avec une agitation de 80 rpm. Le volume de « phage helper » est calculé selon la formule :

$$(5*10^{8}*DO*V_{culture}*\text{facteur de multiplicité}) / (\text{Titre du phage hepler})$$

Les bactéries infectées sont collectées par centrifugation (10 min, 4000g) puis resuspendues dans 20 mL 2YT+Amp+Kan. Ces cultures de 20mL ont été divisées en deux cultures : une première qui est induite à l'IPTG (50µM) et une deuxième non induite. Ces cultures ont été incubées o.n à 30°C. Les phages produits ont été, par la suite, concentrés et purifiés par précipitation au PEG à partir de chaque surnagent de culture. En effet, les surnageants ont été mélangés 1:1 (V/V) avec du PEG NaCl (20 % PEG 8000, 25 M NaCl) pendant 20' à température ambiante. Les phages précipités sont collectés par centrifugation (30', 12000g) puis ils ont été repris dans du TBS et précipités une $2^{ème}$ fois. Les phages ont été repris dans du TBS puis dosés soit par infection de bactéries XL1Blue MRF' soit par mesure de la DO à 269nm.

Le taux de display est calculé selon l'exposition de protéines ayant le Flag-tag à la surface des phages qui vont reconnaître un anticorps anti-Flag-tag. Ce dernier a été immobilisé sur une plaque ELISA à une concentration de 10 µg/mL o.n. à 4°C. La plaque a été par la suite bloquée avec une solution de TBST BSA3% pendant 2h à température ambiante. 100 µL des phages précipités ont été incubés, sur la plaque, en présence de l'anticorps 1h à température ambiante. Les phages non spécifiques sont éliminés par 8 lavages au TBST. Les phages spécifiques, liés aux anticorps anti-Flag-tag immobilisés, ont été élués par ajout de Glycine HCl pH2. Les phages élués sont neutralisés par ajout de Tris HCl pH8 pour être comptés par la suite.

Les comptages de phages ont été réalisés de la sorte : des dilutions de phages sont préparées puis 100µL des dilutions sont incubés 15min à 37°C avec 200 µL de XL1Blue MRF' à DO=0.6. 100 µL des bactéries infectées ont été étalés sur du milieu solide 2YT agar+Amp+Glu. Les colonies, qui ont poussé sur les boîtes, ont été comptées. Ainsi nous avons pu calculer les phages totaux déposés

sur l'anticorps ou Input et les phages fixés par l'anticorps puis élués ou Output. Ainsi, nous avons déterminé le taux de display ($\tau_{display}$) (Tab1).

				$\tau_{display}$
αRep-n4-a	Phages issus d'une culture induite	*Input*	$1.3*10^8$cfu/mL	***3.4*10^{-4}***
		Output	$4.5*10^4$ cfu/mL	
	Phages issus d'une culture non induite	*Input*	$7.5*10^8$ cfu/mL	***1.2*10^{-5}***
		Output	$9*10^3$ cfu/mL	

Tab1 : Tableau récapitulatif des calculs des *Input*, des *Output* ainsi que le taux de *display* des protéines *αRep* sur le phage M13.

Ces résultats montrent que l'induction à l'IPTG (50 µM) diminue, légèrement, la production des phages ($Input_I < Input_{NI}$) mais ça permet d'augmenter le taux de *display*, et par la suite la fraction des phages qui exposent à leurs surfaces les protéines d'intérêt.

Il est important de souligner que le taux de display obtenu ici est très bas. Avec la NCS dans le vecteur *phDiEx* le taux de *display* était plutôt de l'ordre de 10^{-3}, sans induction. Ce taux d'exposition faible peut avoir plusieurs explications. Il peut s'agir d'un problème de mesure du taux d'exposition, par exemple lié à une éventuelle protéolyse du *Flag–tag*. Dans cette hypothèse, l'étiquette accessible serait protéolysée sur une partie de la population de phages pendant la phase de production et cette sous-population n'étant pas comptabilisée conduit à sous-estimer l'efficacité réelle d'exposition. Il est également possible que l'affinité du *Flag-tag*, soit insuffisante et conduise à une dissociation non négligeable de phages retenus sur l'anticorps lors des lavages. Dans ce cas également, le taux d'exposition serait sous-estimé. Enfin, il est possible que le taux d'exposition soit réellement très bas ce qui aurait pour inconvénient d'imposer l'utilisation d'échantillons importants de phages comme *Input* lors des sélections.

Dans cette hypothèse, nous avons cherché, lors des expériences ultérieures, à améliorer le taux d'exposition par utilisation d'un phage auxiliaire décrit (*Phaberge* (Soltes et al. 2003)) pour augmenter le taux de *display*. Ce phage auxiliaire exprime un taux plus faible de protéine III ce qui induit une compétition moins forte avec la fusion *αRep-pIII* codée par le phagemide, et donc à une incorporation plus efficace de la protéine à exposer sur les particules virales en formation. Les évaluations comparées réalisées au laboratoire sur une autre construction suggèrent que ce *phage helper* permet d'augmenter l'efficacité d'exposition d'un ordre de grandeur.

Le degré de protéolyse et/ou de dissociation lors de la mesure sur *Flag-tag* ne sont pas connus précisément. Des mesures ont été récemment réalisées par Anne Chevrel, doctorante au laboratoire, en utilisant non pas le *Flag-tag* pour évaluer le taux de *display* mais l'interaction entre une *αRep* préalablement sélectionnée pour fixer efficacement (Kd de l'ordre du nM) une protéine cible. Dans ce cas, et en utilisant le *phage helper Phaberge* et sans induction à l'IPTG, le taux de *display* mesuré est de l'ordre de 10^{-2}. Ce résultat est nettement plus favorable que ce que nous avions mesuré. Il est peu probable que le changement de phage *helper* soit à lui seul responsable de cette amélioration de trois ordres de grandeurs et il parait donc vraisemblable que la mesure du taux d'exposition en utilisant le *Flag-tag* soit sous-estimée.

4. Construction du vecteur accepteur de la banque d'αRep de 2éme génération

Le vecteur « accepteur » est une construction intermédiaire conçue uniquement pour permettre une procédure de construction efficace de la bibliothèque. Sa structure générale doit comporter les gènes nécessaires à l'expression et l'exposition des protéines sur phages. Ainsi, ce vecteur comporte l'opérateur lactose, le promoteur T7, la séquence d'export *DsbA* ainsi que les étiquettes : le *Flag-tag* de l'extrémité N-terminale de la protéine et le *His-tag* de l'extrémité C-terminale. Il comprend également la séquence 248-406 du gène III du phage M13, le gène de résistance à l'ampicilline ainsi que les sites de restriction essentiels pour la construction de la banque. La partie essentielle de ce vecteur qui permet l'assemblage des gènes d'*αRep* est une cassette qui comporte une partie du *Ncap*, le *Ccap* et entre les deux les sites de restrictions au sein duquel seront intégrés les modules répétés. La cassette « acceptrice » a donc dû être assemblée à partir d'oligonucléotides dans le vecteur destinée à la construction de la banque de seconde génération.

La construction du vecteur « accepteur » a été réalisée à partir du clone αRep-n4-a ayant la nouvelle construction du promoteur et le Flag-tag comme étiquette du coté N-terminal de la protéine. En effet, le clone αRep-n4-a avec pLac-T7prom-Flag-tag et l'ancien vecteur accepteur de la banque αRep de première génération ont été digérés NotI- HindIII (Fig. 14).

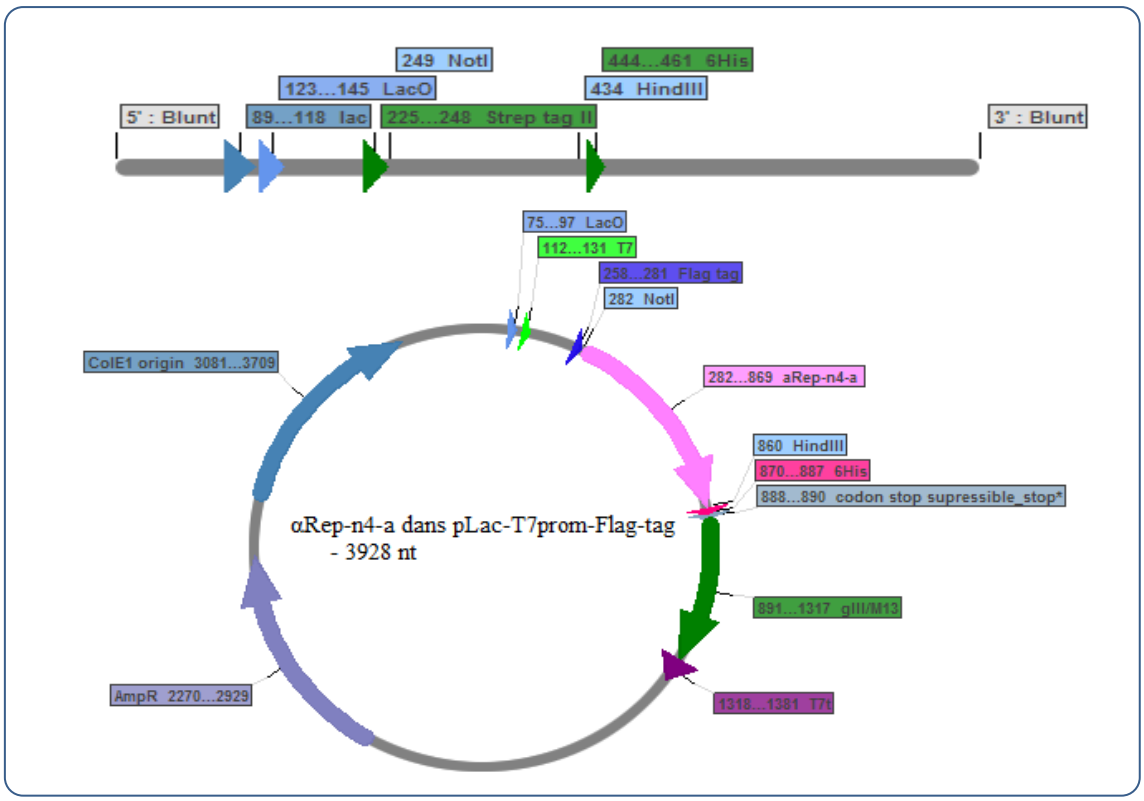


Fig. 14 : Le plasmide codant pour de l'αRep-n4-a avec *pLac-T7prom-Flag-tag* et le module accepteur de la banque de première génération avec les sites de restriction NotI-HindIII utilisés pour le clonage.

Le fragment d'ADN comprenant le nouveau promoteur et l'étiquette Flag a été par la suite cloné dans le vecteur accepteur de la banque de première génération. 9 clones issus du clonage et ayant le bon profil dans la digestion contrôle ont été séquencés en utilisant 7 amorces. Ces 7 amorces ont été conçues de sorte à permettre la vérification de la séquence entière du plasmide (Fig. 15 et 16). Les 7 amorces utilisées sont les suivantes :

P1 (0 pb à 600 pb) : CAATACGCAAACCGCCTCTCC

P2 (500 pb à 1100 pb) : GAACTATCTCGAGACCCATAA

P3 (1000 pb à 1600 pb) : CGTTTGCTAACATACTGCGTAA

P4 (1500 pb à 2100 pb) : TATAAGGGATTTTGCCGATTTC

P5 (2000 pb à 2600 pb) : GGTCGCCGCATACACTATTCTC

P6 (2500 pb à 3100 pb) : CGTAGTTATCTACACGACGGGG

P7 (3000pb à 3600 pb) : GCTGCTGCCAGTGGCGATAAGT

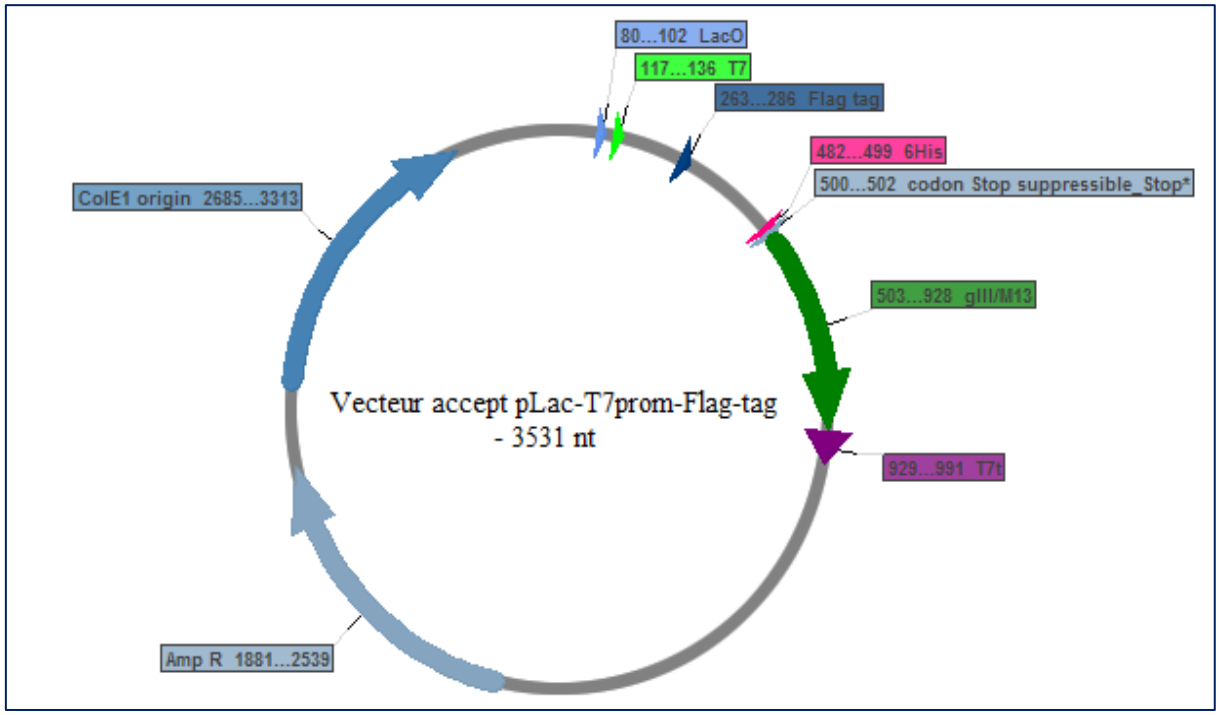


Fig. 15 : Carte du vecteur accepteur de la banque *αRep* montrant les promoteurs Lac O puis T7prom suivie de la séquence d'export, le *Flag-tag* et finalement le *His-tag* qui marque la fin de la séquence insérées.

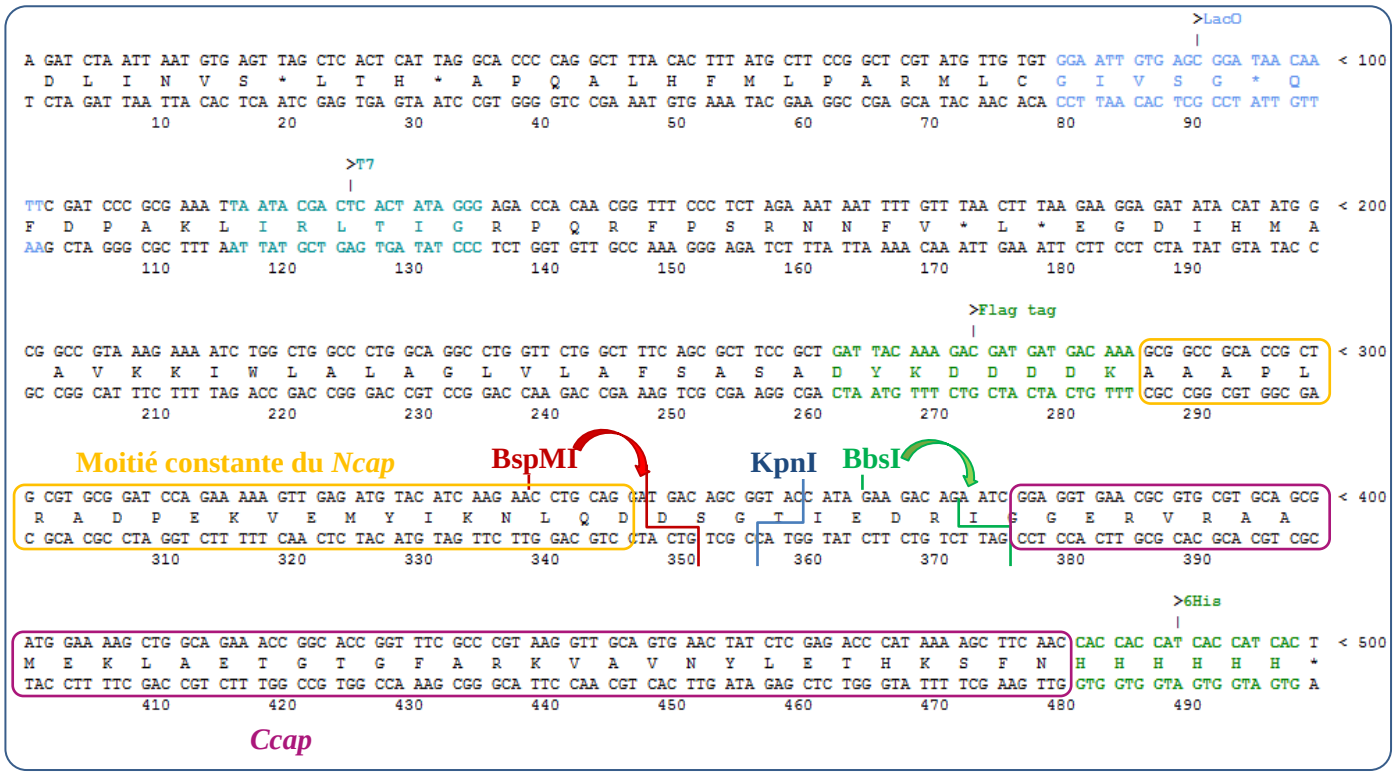


Fig. 16 : Séquence entre 80pb et 499pb du vecteur «accepteur» montrant les gènes Lac O et T7prom suivis de la séquence d'export et le *Flag-tag*, le *Ncap* et le *Ccap* ainsi que les sites de coupure des enzymes de restriction (KpnI, BspMI et BbsI) qui seront utilisées pour la construction de la banque.

Conclusion

Une fois le vecteur « accepteur » construit et la séquence vérifiée, l'étape suivante de notre travail était de construire une nouvelle banque d'*αRep* dite de deuxième génération. Cette banque doit être améliorée par rapport à la banque *αRep* de première génération à deux niveaux :

- Amélioration de la distribution des acides aminés au niveau des positions hypervariables : Dans la première banque, les positions hypervariables ont été codées par des codons partiellement dégénérés, dans le but principal d'évaluer la tolérance des positions à une variabilité de séquence. Ces codons avaient été choisis pour coder des acides aminés aux propriétés variées mais les distributions qui en résultent comportent des biais arbitraires. Cela pouvait être amélioré et dans la bibliothèque de deuxième génération, la distribution des fréquence d'acides aminés a été codée position par position, de façon à tendre vers la distribution des fréquences des acides aminés à chaque position telle qu'elle est observée dans la collection de modules répétés naturels
- Amélioration de la proportion des clones codants : il semble essentiel de parvenir à construire une bibliothèque comportant une proportion plus élevée de modules codants. Cela supposait soit d'éviter autant que possible les erreurs lors de l'assemblage de la bibliothèque, soit à corriger ou éliminer ces erreurs ensuite.

Ces aspects sont traités dans la section suivante des travaux de cette thèse.

B. Optimisation de la fraction des clones codants dans la banque

Introduction

Avant de construire une nouvelle banque, nous nous sommes intéressés à l'amélioration de la qualité de nos banques ce qui revient à déterminer l'origine éventuelle des erreurs de séquence observées dans les variants non-codants.

Des erreurs sont toujours présentes lors de synthèse d'oligonucléotides, et pour en minimiser l'occurrence, il est préférable de purifier les oligonucléotides, ou encore de se limiter à utiliser des oligonucléotides de taille modérée (typiquement < 60 nt) d'autant plus que la population de molécules sans erreurs décroit avec la taille des oligonucléotides synthétisés. Mais la procédure de construction de la bibliothèque comporte aussi des possibles sources d'erreurs spécifiques à la procédure que nous avons utilisée. Par exemple les erreurs peuvent provenir d'appariements incorrects des oligonucléotides lors de l'assemblage des cercles ou du processus d'amplification circulaire utilisés (RCA) (Fig. 1).

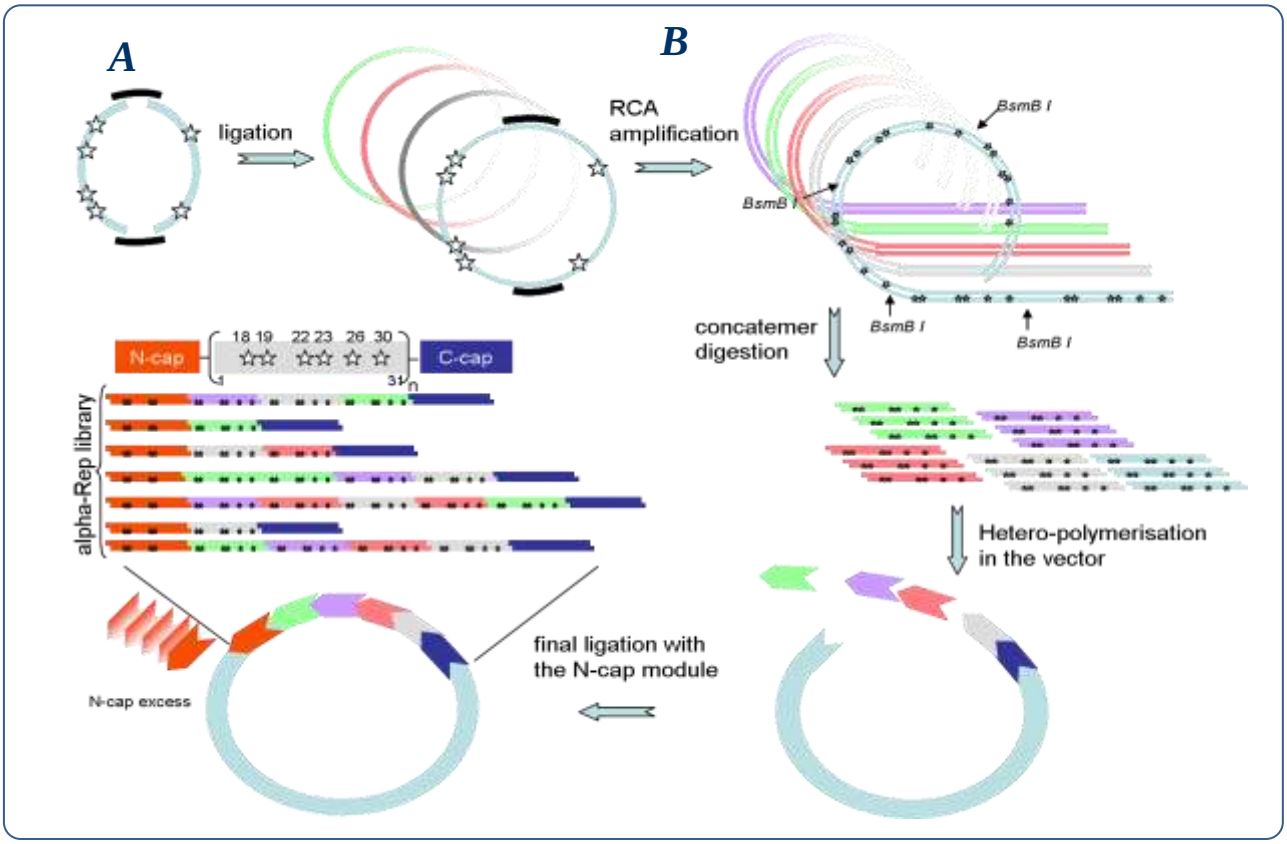


Fig. 1 : Schéma général de la construction de la banque d'*αRep* : A : Première hypothèse : les erreurs de séquences proviennent de l'hybridation des oligonucléotides simples brins lors de la formation des cercles. B : Deuxième hypothèse : les erreurs de séquences proviennent du processus d'amplification (RCA).

1. Les erreurs de séquences : quelle origine?

L'une de nos hypothèses de travail était que les mutations détectées dans les séquences des clones de la banque de première génération (erreurs) pouvaient provenir des amorces aléatoires qui sont utilisées lors de la procédure d'amplification circulaire (RCA). En effet, le principe de cette amplification isotherme est que toute chaîne en cours d'élongation peut servir elle-même de matrice. Le mélange réactionnel comporte à cette fin des petits oligonucléotides de séquences aléatoirement variables. Par exemple, le kit d'amplification que nous avons utilisé *TempliPhi (GE)* comporte un tampon de réaction (Reaction Buffer) qui contient des hexamères aléatoires qui vont jouer le rôle d'amorces pour l'amplification. Toute séquence d'ADN simple brin apparaissant dans le milieu réactionnel du fait de l'activité de déplacement de brin de la polymérase, peut alors s'hybrider aux hexamères aléatoires qui lui sont complémentaires. Or, pour l'amplification d'une séquence répétitive dont l'unité répétée est de petite taille (93 nucléotides) les hexamères qui se trouveront à s'hybrider avec les régions répétées seront très fréquemment utilisés et par conséquent rapidement consommés par l'amplification, tandis que d'autres ne correspondent à aucune région de l'ADN matrice, persisteront dans le mélange réactionnel. Il paraissait ainsi possible que cela puisse engendrer rapidement un fort biais de composition des hexamères aléatoires lors de l'amplification. Cela pourrait favoriser les hybridations imparfaites avec les hexamères restants et donc conduire à des erreurs dans une proportion importante des séquences amplifiées.

1.1. Les amplifications réalisées

Nous avons vérifié la possibilité d'amplifier nos plasmides par le biais de primers dont la séquence est prédéfinie. Des essais d'amplification de cercles d'un motif dont la séquence correspond à la séquence du motif consensus de la banque, ont été réalisés :

- Amplification des cercles par un mélange réactionnel ne contenant pas d'hexamères aléatoires mais à la place un mélange de six oligonucléotides amorces conçus de façon à ce que leurs séquences correspondent aux régions conservées de la séquence répétée. Ces oligonucléotides sont synthétisés avec des extrémités protégées par un groupement phosphorothioate évitant ainsi toute dégradation par l'activité exonucléasique de la polymérase
- Amplification des cercles par le kit *TempliPhi* contenant les hexamères aléatoires et la Φ29 polymérase.

Dans un premier test, nous nous nous sommes proposés d'amplifier des cercles ayant une même séquence avec deux sortes d'amorces : des amorces de séquences déterminées et d'autres de séquences aléatoires permettant ainsi de préciser l'origine de l'erreur.
Quatre amplifications ont été réalisées indépendamment : 3 amplifications contenaient 3 concentrations différentes des 6 amorces conçues pour reconnaître les régions constantes du consensus (0.1 µM, 0.5 µM et 1 µM par primer) et une amplification par le kit *TempliPhi*™ de *GE* contenant les hexamères aléatoires.

- ***Amplification des cercles du motif consensus avec les amorces synthétisés: « Amplification home made »***

Un mélange contenant les cercles du motif consensus, le tampon de la Φ29, les dNTP et les primers synthétiques a été préparé. Ce mélange a été tout d'abord chauffé 3 min à 95°C puis refroidi jusqu'à 4°C pour permettre une bonne hybridation des amorces de petite taille (hexamères). Les enzymes nécessaires à la polymérisation: 5 unités de Φ29 polymérase et la pyrophosphatase ainsi que la BSA ont été par la suite ajoutées.

- ***Amplification des cercles du motif consensus avec le kit TempliPhi™ de GE:***

Pour cette amplification, les cercles ont été mélangés avec le « Sample Buffer » du kit. Ce tampon contient des hexamères aléatoires non spécifiques servant d'amorces pour l'amplification. Le mélange a été ensuite chauffé à 95°C pendant 3 min permettant la déshybridation des cercles doubles brins. Par la suite, le refroidissement du mélange à 4°C servira à hybrider les hexamères avec l'ADN matrice. La polymérisation a nécessité l'ajout de l'«Enzyme mix», contenant la Φ29 polymérase et des hexamères aléatoires puis ajout du «Reaction buffer», qui contient les sels et les deoxynucléotides et ajuste le pH optimal à la synthèse de l'ADN.

Les deux types de mélanges ont été incubés pendant 4h et o.n. à 30°C pour la synthèse de l'ADN. Les différents produits de RCA sont par la suite repris dans 240 µL d'eau pour arrêter la réaction. L'ADN synthétisés est digéré avec l'enzyme BsmBI, qui coupe les polymères synthétisés en motifs de 100 pb, afin de vérifier l'efficacité de l'amplification. En effet, pour chaque 135 µL de produit de RCA, on a jouté soit 0.2 µL de BsmBI soit 1 µL de BsmBI et 15 µL de tampon NEB3. Ces échantillons ont été incubés respectivement 5 min et 1h 30 min à 55°C et les digestions sont par la suite purifiées et déposées sur un gel d'agarose 2% (Fig. 2).

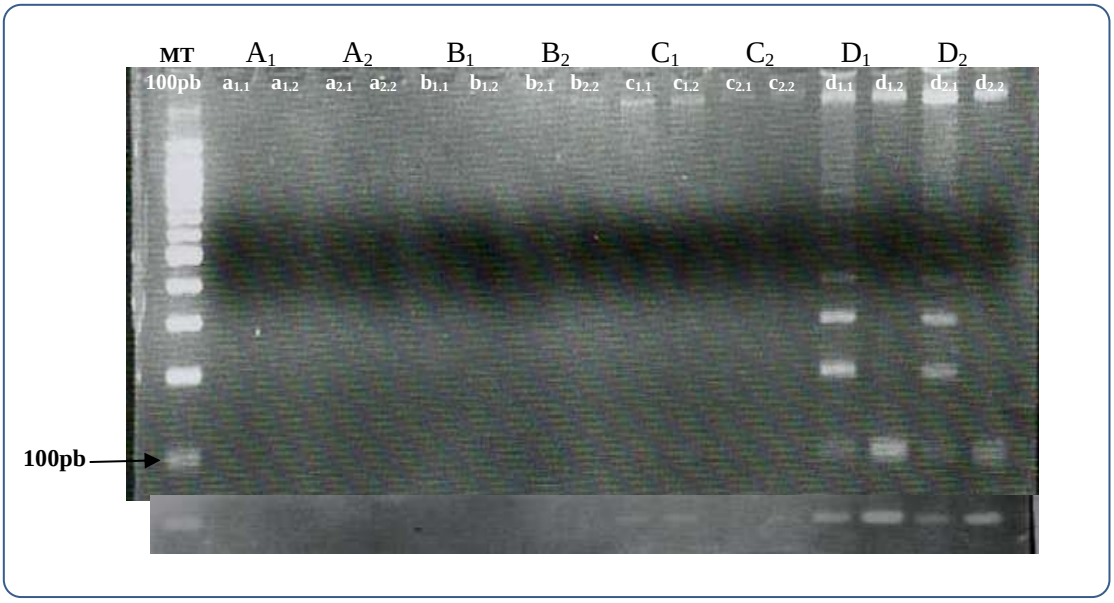


Fig. 2 : Séparation des produits obtenus après digestion BsmBI des produits issus des différentes amplifications des cercles du motif consensus.
A: Amplification « *home made* » avec une concentration de 0.1 µM par primer avec une incubation de 4h A1et o.n. A2.
B : Amplification « *home made* » avec une concentration de 0.5 µM par primer avec une incubation de 4h B1 et o.n. B2.
C : Amplification « *home made* » avec une concentration de 1 µM par primer avec une incubation de 4h C1 et o.n. C2.
D : Amplification avec le Mix commercial avec le kit illustra *TempliPhi*™ de *GE* avec une durée de 4h D1 et o.n. D2.
La différence entre a1.1, b1.1, c1.1 et d1.1 avec a1.2, b1.2, c1.2 et d1.2 est le type de digestion : les échantillons notés .1 correspondent à une digestion partielle alors que les échantillons notés .2 correspondent à une digestion totale.

Les résultats obtenus montrent une bande proche de 100 bp correspondant aux polymères résultants de l'amplification coupés par l'enzyme BsmBI dont les sites sont situés entre chaque motif répété. La concentration optimale des amorces spécifiques synthétisées est de 1µM (par primer). Il faut toutefois noter que le kit commercial reste toujours plus efficace que le mélange « *home made* ». Nous avons aussi constaté que les 4 heures de synthèse d'ADN permettent d'avoir une quantité de motifs amplifiés plus importante que les synthèses sur une nuit. L'objectif suivant a été de vérifier si ces cercles dont la séquence est préalablement connue sont soumis à des mutations lors de leur amplification et si la fréquence de ces mutations dépend des oligonucléotides utilisés lors de l'amplification.

1.2. Vérification des séquences des motifs amplifiés

Dans le but de vérifier la fidélité de la polymérisation, l'amplification par RCA a été utilisée pour amplifier un module unique d'*αRep* de séquence définie correspondant à la

séquence du motif consensus y compris aux positions variables. La séquence codant ce module a été extraite d'un clone comportant un gène consensus préalablement séquencé. La séquence correspondante est isolée par restriction, purifiée puis liguée pour être utilisée comme ADN matrice pour la réaction d'amplification. L'ADN amplifié a été par la suite utilisé pour construire une mini-bibliothèque test. Le séquençage d'un ensemble de clones pris au hasard au sein de cette bibliothèque permet de vérifier s'ils sont tous constitués de la répétition du même module, exactement identique au module consensus amplifié, ou à l'inverse si des mutations ont été introduites lors de l'amplification.

Au sein du laboratoire, nous disposons d'une banque «*αRep* homorepeat » constituée de variants qui différent les uns des autres par le nombre du même motif inséré. Nous avons choisi un variant de cette banque contenant un seul motif inséré : **HR-n-1** dans lequel les motifs amplifiés par RCA seront insérés.

En effet, ce clone est tout d'abords digéré BsmBI-BsaI puis purifié. Les motifs amplifiés par le kit commercial TempliPhi™ (voir paragraphe précédent) sont par la suite ajoutés au clone HR-n-1 digéré puis ligués en présence de T4 DNA ligase o.n. à 16°C. Le produit de ligation a été purifié puis utilisé pour électroporer des XL1Blue MRF' permettant ainsi d'obtenir une mini-banque d'αRep d'«homorepeat». 20 clones indépendants ont été pris au hasard de cette mini-banque, leurs plasmides ont été extraits et séquencés.

Les clones obtenus contiennent entre 1 et 9 modules insérés. L'alignement des différents motifs insérés dans ces clones (au nombre de 56) est réalisé par le logiciel d'alignement en ligne «*Multiple sequence alignment with hierarchical clustering*» sur le site *ExPaSy tools*. L'alignement a montré très clairement qu'aucune erreur n'a été introduite au cours de l'amplification. En effet, sur 20 clones soit 56 modules séquencés aucune erreur de séquence n'a été introduite.

La fréquence d'erreurs induite par l'amplification est donc basse et très inférieure à la fréquence des mutations observées dans la bibliothèque de première génération. Il apparait donc que les erreurs ne proviennent pas, dans leur grande majorité, du processus d'amplification : ni de la polymérase utilisée (la Φ29 polymérase) ni des amorces aléatoires comme nous l'avions envisagé.

Cela signifie que les erreurs observées dans la banque ont pour origine non pas le processus d'amplification mais plutôt les oligonucléotides synthétiques eux-mêmes ou

d'éventuels mésappariements lors de la constitution des cercles utilisés comme matrice lors de l'étape d'amplification par RCA. Une conclusion annexe, mais importante, est que si l'amplification n'est pas ou peu mutagène, alors il est concevable de pouvoir amplifier une population de modules de séquences correctes sans que l'amplification n'y introduise de mutations nouvelles. En d'autres termes, il parait possible de sélectionner le sous ensemble de la banque dont les protéines ne comportent que des modules corrects, d'extraire ceux-ci pour les réamplifier sans perte de qualité, puis de les recombiner.

Une telle stratégie n'aurait pas été envisageable si l'amplification par RCA avait été la source principale des erreurs, puisque une population de modules dont les erreurs sont éliminées par « filtration » redeviendrait incorrecte dès lors qu'elle serait à nouveau amplifiée.

A ce stade, il devenait donc concevable, et important, de parvenir à sélectionner la population de clones effectivement codants et pour cela de mettre au point une méthode efficace de « filtration » par exposition sur phages.

2. Essais de filtration des phages

Nous avons alors décidé de tester puis d'optimiser une méthode reposant sur la sélection par *phage display* de protéines dont l'étiquette *Flag* est accessible, signe que l'extrémité N-terminale de la séquence est en phase avec le domaine d'ancrage de la protéine sur le phage. Nous avons tout d'abord testé des procédures de filtration des phages avant de les appliquer pour la construction de la banque de deuxième génération.

Pour ces essais préliminaires, nous avons opté pour une stratégie de filtration à partir de mélanges en proportion prédéfinies de deux populations clonales de phages correspondant à un phagemide comportant une séquence correcte d'*αRep* et un autre ne comportant pas de séquence codante en phase avec la protéine pIII du phage. Pour mimer une sélection nous avons réalisés des mélanges comportant une proportion minoritaire : 1% et 10 % de phages codants parmi une majorité de phages non codants. Lors de ces essais nous avons également comparés les résultats obtenus avec différentes procédures de préparation de phages à « filtrer » ainsi qu'avec deux types de phages *helper*.

L'exposition de l'*αRep* sur phage a été réalisée par infection des bactéries par le phage *helper* M13KO7 d'une part et son dérivé nommé *Phaberge* d'autre part (Soltes et *al.*, 2003). L'utilisation de l'un ou de l'autre phage *helper* vise à déterminer celui qui permettait d'augmenter le taux de *display* et par la suite améliorer le processus de filtration. En effet, le

phage auxiliaire *Phaberge* a été obtenu en introduisant une mutation qui transforme le résidu Q350 en un codon stop, suppressible dans certaines souches bactérienne (souches supE par exemple *XL1 Blue MRF'*). Cette mutation est localisée dans le gène de la protéine pIII du phage *helper* M13KO7 (Vieira and Messing, 1987).

Nous avons également réalisé des essais exploratoires portant sur la procédure de purification des phages. En effet, l'observation par microscopie électronique d'échantillons de phages préalablement précipités au PEG montre invariablement qu'une fraction des phages demeurent accolés les uns aux autres et n'ont pas été complétement séparés après leur agglomération par le PEG (Fig. 3).

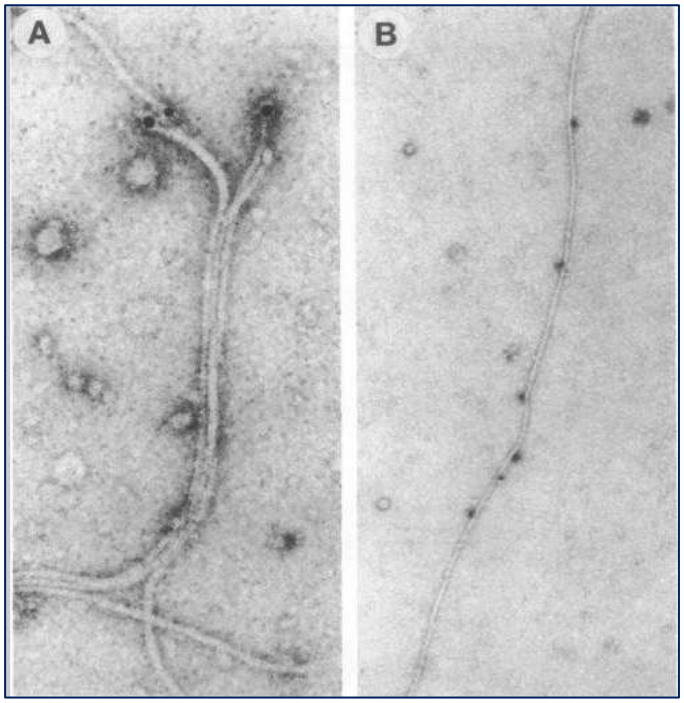


Fig. 3 : Images en microscopie électronique de Bactériophage exposant des domaines anticorps sur la protéine III (A) ou VIII (B). Les phages accolés sont visibles en A (Barbas et *al.*, 1991).

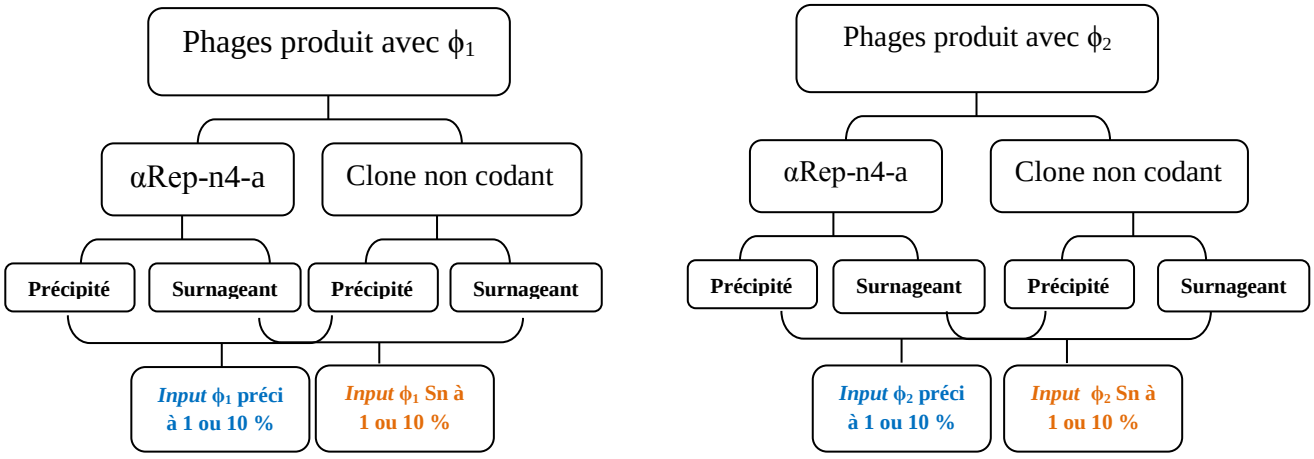
La formation de ces « fagots » de phages peut potentiellement perturber la sélection puisqu'un groupe de phages comportant des séquences incorrectes peut être retenu sur un anticorps anti-*Flag-tag* si seulement une des particules du groupe comporte l'étiquette. Les autres particules non correctes sont ainsi sélectionnables de façon illégitime grâce à leur association et vont être finalement propagées si l'une des particules incorrectes parvient à infecter une bactérie réceptrice. La présence de phages agglomérés n'est pas nécessairement incompatible avec une sélection efficace si les cycles de sélection se répètent, mais elle pourrait gravement perturber une étape unique de sélection de la sous population codante.

Ainsi, nous avons décidé de comparer différentes stratégies pour la préparation des phages en utilisant soit des phages précipités au PEG, soit des phages prélevés des surnageants de cultures. Et nous avons fait des essais avec des mélanges prédéfinis de phages qui sont préparés de deux manières différentes : précipités ou prélevés directement du surnageant des cultures

2.1. Production des phages, préparation des *Input* pour la filtration

Dans le but de mettre au point le procédé de « filtration », nous nous sommes basés sur une interaction αRep/αRep identifiée et caractérisée par des travaux préalables du laboratoire. L'αRep-n4-a peut interagir avec une autre *αRep2E3* avec une constante de dissociation de quelques nanomolaire. Des phages, exposant l'αRep-n4-a, ont été mélangés à proportions minoritaires avec des phages non-codants puis mis en contact avec l'*αRep2E3* purifiée et fixée sur plaque. Par la suite, le ratio entre le nombre de phages ajoutés sur plaque (*Input*) et le nombre de phages élués après fixation (*Output*), va nous permettre de déterminer l'efficacité de la filtration.

Deux cultures de phages ont été préparées comme décrit dans A-III-3 à partir de bactéries exprimant l'αRep-n4-a dans phDiEx et d'autres exprimant un clone non-codant. D'une part, deux types de phages helper ont été utilisés: le M13KO7 (Φ1) et le Phaberge (Φ2). Et d'autre part, les Input ont été préparés de deux façons différentes : un mélange de phages précipités au PEG et un mélange de phages pris directement des surnageants de cultures comme présenté sur le graphe suivants:



Sur le plan pratique, l'αRep 2E3 a été immobilisée, sur plaque ELISA, à une concentration de 20 µg/mL o.n. à 4°C et 400rpm. Par la suite, la plaque a été bloquée avec une solution de TBST BSA3%. Les 4 types d'Input ont été incubés 1h30 à 25°C. Les phages non-spécifiques sont éliminés par 20 lavages au TBST et 20 lavages au TBS. Les phages fixés spécifiquement à l'αRep 2E3 ont été libérés par élution acide.10 µL des phages élués ont été utilisés pour le comptage et le reste a été utilisé pour infecter, indépendamment, 5mL de bactéries XL1Blue MRF'.

2.2. Analyse des résultats de la filtration : Comparaison entre *Input* et *Output*

Nous avons choisie dans cet essai de partir de deux clones préalablement caractérisés : séquences et propriétés d'expression connues. L'efficacité de la filtration peut être évaluée alors par digestion enzymatique des plasmides des deux clones ou encore grâce aux propriétés d'expression distinctes de ces deux variants.

- Résultats de la filtration par digestion enzymatique des plasmides

L'efficacité de la filtration a été analysée par digestion NdeI-HindIII. En effet, la digestion NdeI-HindIII du plasmide du clone non-codant et de celui de l'αRep-n4-a donne des profils complètement différents (Fig. 4).

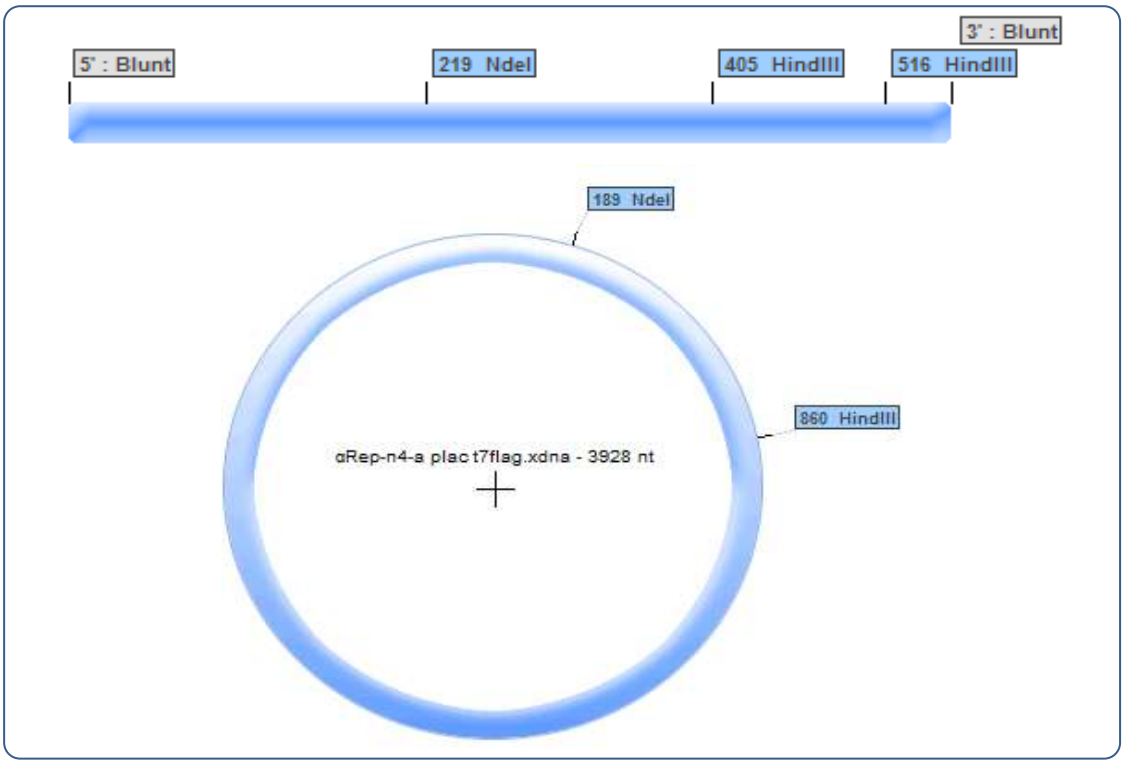


Fig. 4 : Les sites de restrictions de NdeI et HindIII sur les plasmides du clone non-codant et de l'αRep-n4-a : La digestion NdeI-HindIII du plasmide du clone non-codant génère 2 bandes (186 pb et 111 pb) alors que la même digestion du clone αRep-n4-a génère une seule bande de 671 pb.

Nous avons alors extrait les plasmides des bactéries représentatives des 8 Output obtenus. Par la suite, des échantillons de même quantité d'ADN de chaque Output ont été digérés NdeI-HindIII puis déposés sur gel agarose 1% (Fig. 4).

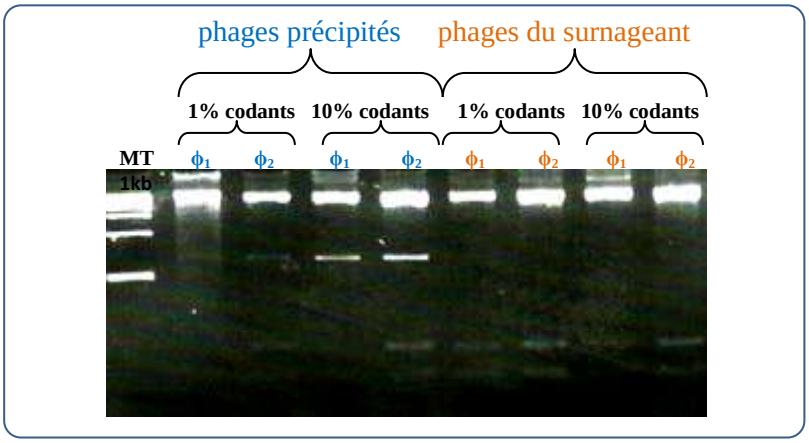


Fig. 5 : Profil d'électrophorèse des digestions NdeI-HindII des différents ADN des *Output* de la filtration de l'αRep-n4-a par l'*αRep* 2E3.

Ces profils montrent que la filtration est meilleure quand l'*Input* contient 10% de codants et qu'elle n'est efficace que pour les échantillons de phages précipités et non, contrairement à ce que nous avions envisagés, à partir de ceux prélevés directement des surnageants de cultures.

Par comparaison entre les filtrations des phages issus des cultures infectée par le phage *helper* M13KO7 (Φ1) et ceux issus des cultures infectées par le *Phaberge* (Φ2), nous remarquons que *Phaberge* permettait une meilleure filtration.

Toutefois aucune condition ne permettait d'éliminer complètement les phages non-codants.

- Analyse de l'efficacité de la filtration par le biais des propriétés d'expression des deux clones testés

Les résultats de la filtration peuvent être aussi analysés en se basant sur les propriétés d'expression des clones utilisés. Au sein du laboratoire, nous disposons d'un test d'expression en milieu solide : *Cofiblot*. Ce test permet de tester simultanément l'expression de 72 clones différents par plaque, par comparaison à l'expression d'un témoin positif et un autre négatif. On peut ainsi estimer le pourcentage des clones codants issus de la filtration et par la suite évaluer l'efficacité de la procédure utilisée pour cela.

Lors du *Cofiblot*, des colonies de bactéries, exprimant les variants à tester, sont cultivées sur un filtre *Durapore* déposé sur milieu solide. Le filtre est, après croissance des colonies, transféré sur milieu solide contenant l'inducteur de l'expression (l'IPTG). Les bactéries sur le filtre expriment ainsi la protéine si elles comportent une séquence codante. Elles sont alors lysées et les protéines, exprimées sous forme soluble, sont transférées sur membrane de nitrocellulose puis révélées par un anticorps anti-*His-tag*. Si la protéine est codante le *tag* sera alors reconnu par l'anticorps et donnera un signal positif alors que les clones non-codants ne seront pas détectés.

Les différents échantillons d'ADN correspondant aux Output ont été utilisés pour transformer, indépendamment, des bactéries Rosetta Blue. Par la suite, 4 plaques 96 puits de culture ont été préparées selon le même schéma représenté sur la Fig. 6 : Ensemencement d'une colonie par puits pour les puits de 1-12 de A à F. La ligne G (puits G1, G3 et G5) est réservée au témoin positif (αRep-n4-a dans Rosetta Blue) et la ligne H (les puits H2, H4 et H6) au témoin négatif (Clone non-codant dans Rosetta Blue).

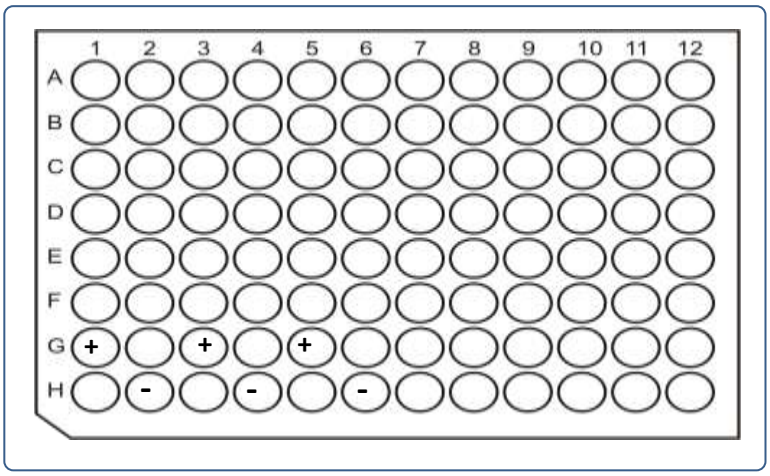


Fig. 6 : Plan d'une plaque 96 puits pour un *Cofiblot* : les puits de A1 jusqu'à F12 sont ensemencés avec des bactéries transformées par les plasmides des *Output*. Les puits G1, G3 et G5 : témoin positif αRep-n4-a dans *Rosetta Blue* et les puits H2, H4 et H6 : témoin négatif clone non-codant dans *Rosetta Blue*.

La plaque 96 puits a été incubée o.n. à 37°C à 500 rpm puis répliquée sur un filtre Durapore sur milieu solide 2YTagar+Amp o.n. à 37°C. Par la suite, le même filtre a été incubé sur milieu solide 2YTagar+Amp+IPTG permettant ainsi l'induction de l'expression des protéines. Après 4h d'induction, le filtre Durapore a été déposé, colonies vers le haut, sur un « sandwich » composé d'un papier Wattman et d'une membrane de nitrocellulose. Le sandwich a été imbibé de 5 mL de tampon de lyse B-PERII permettant ainsi la lyse des bactéries et la libération des protéines. Seules les protéines solubles traversent le filtre par capillarité et vont s'imprégner dans la membrane de nitrocellulose.

Cette dernière a été finalement révélée par un anticorps anti-His-tag couplé à la peroxydase. Les colonies exprimant le variant codant sont détectées par l'apparition de la coloration bleue qui résulte de la réaction de la peroxydase de l'anticorps avec le substrat précipité ajouté (Fig. 7)

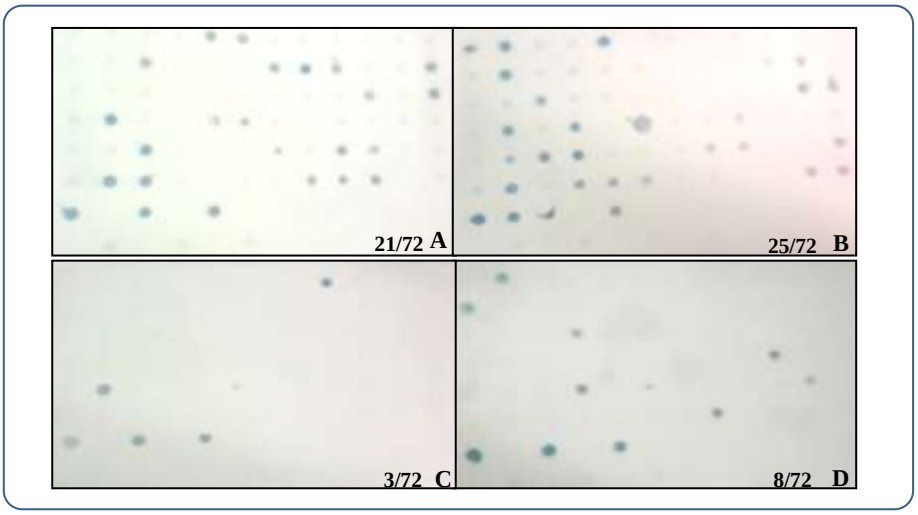


Fig. 7 : Membranes du *Cofiblot* révélées correspondant à : A : *Output* de la filtration dont l'*Input* est précipité à 1% de codants produits avec le phage *helper* ϕ_1. B : *Output* de la filtration dont l'*Input* est précipité à 1% de codants produits avec le phage *helper* ϕ_2. C : *Output* de la filtration dont l'*Input* est prélevé du surnageant de culture à 1% de codants produits avec le phage *helper* ϕ_1. D: *Output* de la filtration dont l'*Input* est prélevé du surnageant de culture à 1% de codants produits avec le phage *helper* ϕ_2.

La filtration des phages exposant l'*αRep-n4-a* avec l'*αRep*2E3 a permis d'augmenter le pourcentage de codants de 1% à 29% quand les phages de l'*Input* sont produits avec le *M13KO7* et peut atteindre 35% quand les phages sont produits avec le phage auxiliaire *Phaberge*. La filtration s'avère moins efficace lorsque les phages sont prélevés des surnageants de cultures : on atteint les 10% de codants avec *Phaberge* et que 4% avec *M13KO7*.

Les résultats du *Cofiblot* confirment ce qui a été observé avec la digestion des plasmides. En effet, la filtration des phages issus du *Phaberge* (Fig. 7.A et C) est plus efficace que celle réalisée avec les phages issus du phage *M13KO7* (Fig. 7.B et D). L'efficacité de la filtration est aussi dépendante de la préparation des phages : les phages précipités sont filtrés avec une efficacité meilleure que celle des phages prélevés des surnageants de cultures. Toutefois, ces conditions ne permettent en aucun cas d'éliminer totalement les phages non-codants.

Suite à ces observations, nous avons décidé de trouver une stratégie qui permettait de purifier les phages du surnageant des cultures sans avoir recourt à la précipitation.

2.3. Essai de filtration avec les phages dialysés

La filtration est, de façon étonnante, très peu efficace si l'on utilise comme *Input* les phages contenus dans le surnageant. Nous avons envisagé que cela pourrait être dû à une éventuelle compétition entre la protéine libre dans le surnageant de culture et celle qui est exposée à la surface du phage. Si la protéine libre est secrétée dans le surnageant de culture, sous forme non associée aux phages, elle peut probablement entrer en compétition avec celle qui est exposée sur phage dans la reconnaissance de la protéine immobilisée.

Le problème est alors de parvenir à purifier les phages du surnageant de culture sans faire appel à une procédure de purification au PEG. Nous avons pensé utiliser pour cela une étape de dialyse. En effet, les surnageants de culture peuvent être dialysés par le biais d'une membrane qui soit perméable aux petites molécules composant le milieu de culture ainsi qu'aux protéines sécrétées si la porosité des membranes de dialyse est assez large.

La dialyse du surnageant de culture, par une membrane *Spectra/Por® Biotech Cellulose Ester* (CE) de MWCO de 300kDa, permet de garder les phages à l'intérieur de la membrane qui est potentiellement perméable aux protéines sécrétées dans le milieu de culture sous forme non associée aux phages. La dialyse devait alors permettre de purifier les phages sans les agglomérer et améliorer ainsi l'efficacité de la filtration.

Pour tester cette procédure, nous avons reproduits des phages codants et d'autres non-codants. Ils ont été dialysés et par la suite mélangés en proportion de 10%. Ces populations mixtes ont été alors soumises à un essai de filtration de l'αRep-n4-a avec l'*αRep*2E3. Une comparaison a été effectuée entre *Input* dialysé, précipité ou prélevé du surnageant de culture contenant chacun la même proportion de 10% de clones codants.

Comme pour le test précédent, les *Output* obtenus ont été analysés à leur tour par digestion NdeI-HindIII et par *Cofiblot* (Fig. 8).

A partir de ces membranes de *Cofiblot*, nous pouvons estimer que la filtration des clones codants nous a permis d'atteindre un pourcentage de similaire de codants pour l'*Input* dialysé et pour l'*Input* précipité est nettement supérieur que pour l'*Input* prélevé du surnageant.

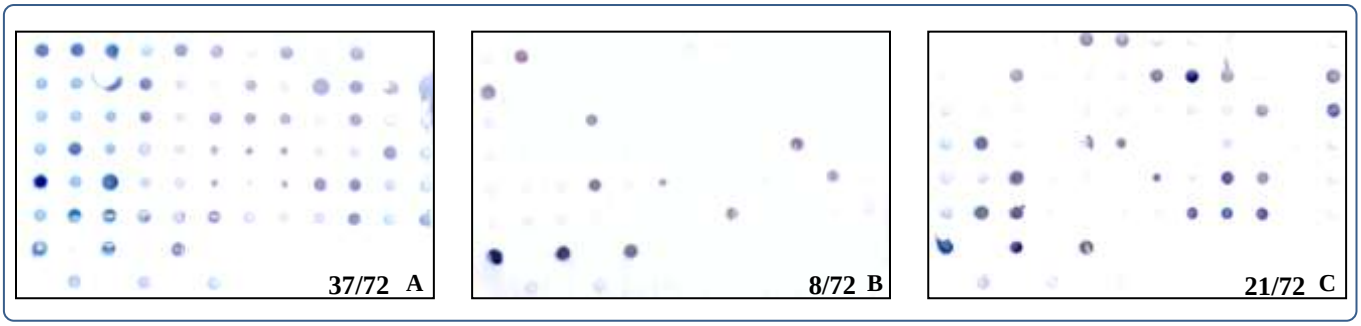


Fig. 8 : Membrane de nitrocellulose comparant les *Output* de la filtration de l'*αRep-n4-a* par *αRep2E3* selon le type d'*Input* : A : *Output* correspondant à l'*Input* de phages dialysés. B : *Output* correspondant à l'*Input* de phages issus des surnageants de cultures. C : *Output* correspondant à l'*Input* de phages précipités.

Il apparait clairement que le traitement par dialyse des phages conduit à une sélection plus efficace que l'utilisation des surnageants de culture. La comparaison entre phages dialysés et phages précipités est plus incertaine puisque le bruit de fond de l'expérience est plus élevé, mais l'efficacité de la sélection semble au moins égale pour les deux types de traitement (phages dialysés et précipités).

Par ces tests, nous avons pu vérifier qu'il est possible d'enrichir la population de clones codants grâce à une interaction spécifique et tester les procédures à utiliser pour réaliser cet enrichissement.

2.4. Optimisation de la concentration de l'anticorps anti-*Flag-tag* utilisé pour la filtration

Nous avons jusque-là mis au point une stratégie globale qui consiste à construire une banque primaire à l'aide de gènes synthétiques puis filtrer les clones non-codants. Dans le but de filtrer ces clones par *phage display*, nous avons décidé d'utiliser l'étiquette *Flag* de l'extrémité N-terminale de la protéine. En effet, les phages, exposant à leur surface les variants codants de la banque primaire, vont reconnaitre l'anticorps anti-*Flag-tag* immobilisé sur des immunotubes. L'objectif étant d'obtenir en un seul passage une population de phages assez diverses pour échantillonner tous les clones codants, il est important de maximiser la capacité de capture de la procédure suivie puis de vérifier qu'elle est suffisante pour retenir le nombre de phages nécessaire. Pour cela, la première étape est d'optimiser la concentration d'anticorps anti-*Flag-tag* à immobiliser dans les immunotubes permettant une filtration efficace des phages codants de la banque primaire d'*αRep*.

Dans le but de déterminer la concentration optimale pour la filtration, nous avons immobilisé différentes concentrations d'anticorps anti-*Flag-tag* sur plaque 96 puits (Fig. 9).

Par la suite, nous avons testé la capacité des anticorps immobilisés à fixer des phages qui exposent une *αRep* avec un *Flag-tag*. Ces derniers sont révélés par des anticorps anti-M13couplé à la peroxydase. Le puits ayant le meilleur signal correspondrait à la concentration qui permettrait de capturer le plus grand nombre de phages exposant une *αRep.*

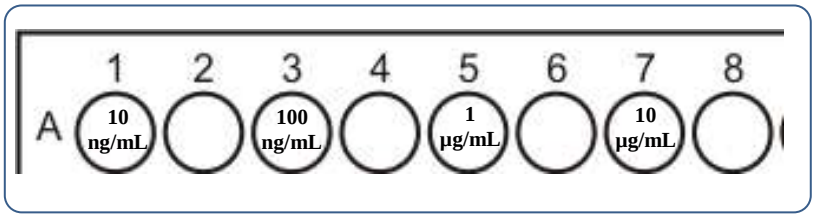


Fig. 9: Plan des concentrations d'anticorps anti-*Flag-tag* dans la plaque : 10ng/mL, 100ng/mL, 1µg/mL et 10 µg/mL.

Sur le plan pratique, nous avons préparé 4 concentrations d'anticorps anti-Flag-tag dans du tampon phosphate (100 µL) : 10 ng/mL, 100 ng/mL, 1µg/mL et 10 µg/mL. Chaque concentration a été incubée dans un puits d'une plaque ELISA pendant 2h à 25°C et 500 rpm. Les anticorps non immobilisés ont été éliminés par 3 lavages au TBST. Les puits sont bloqués ensuite, 2h à 25°C à 500rpm, par 100 µL d'une solution de TBST BSA 3%. Après 3 lavages au TBST, 150µL de phages dialysés, titrés à 10^{12} phages/mL et exprimant à leurs surfaces l'αRep-n4-a, sont mis en contact de l'anticorps 1h à 25°C. Les phages non spécifiques, ne reconnaissant pas l'anticorps, sont éliminés par 10 lavages TBST suivis de 10 lavages TBS. Alors que ceux qui sont en interaction avec l'anticorps anti-Flag-tag sont révélés par le biais d'un anticorps anti-M13-HRP. Cet anticorps est incubé à une dilution de 1/5000, 1h à 25°C. Après 3 lavages au TBST, la présence des phages est révélée grâce au substrat soluble de la peroxydase (Fig. 10). Il est important de signaler que les puits adjacents aux puits contenant l'anticorps ont aussi été bloqués et mis en contact avec les phages dans le but de tester la spécificité de la reconnaissance.

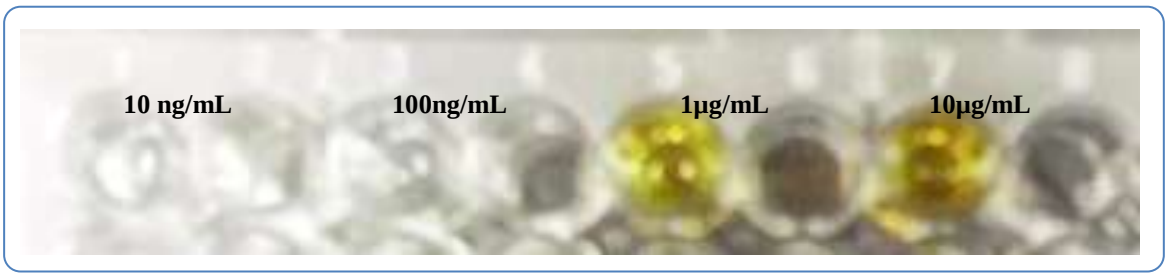


Fig. 10 : Révélation des phages exposant l'*αRep-n4-a* et reconnaissant l'anticorps anti-*Flag-tag* immobilisé dans la plaque ELISA à différentes concentrations.

La coloration jaune qui apparait suite à la réaction entre le substrat soluble et la peroxydase de l'anti-M13 témoigne de la présence des phages et rend compte de la concentration d'anti-*Flag-tag* nécessaire pour leur capture. Ce test Elisa nous a permis d'une part de fixer la concentration optimale à l'immobilisation de l'anticorps dans les immunotubes

pour la « filtration » de la banque primaire à **10 µg/mL**. Et d'autre part, nous avons pu vérifier que cette reconnaissance est spécifique puisque aucun signal n'est détecté dans les puits non-immobilisés.

La stratégie de la filtration a pour but d'éliminer les clones non-codants. Ces derniers ont été estimés de l'ordre de 30% dans la banque de première génération. Une fraction importante des clones longs (qui comportent un nombre important de motifs insérés) sont non-codants puisque plus un clone va comporter de motifs, plus il a de chance d'avoir intégré au moins un motif avec une erreur de séquence. Ainsi, la diversité de la banque primaire en particulier la diversité de taille des protéines est vraisemblablement réduite. Le sous ensemble de la banque primaire sélectionnable par filtration est donc aussi réduit et il sera donc important de parvenir à recréer efficacement de la diversité à partir de la banque filtrée. Nous avons alors pensé à extraire les motifs codants de la banque filtrée et les mélanger de sorte à recréer de la diversité dans la banque finale : une sorte de « *DNA Shuffling* ». Dans le but de mettre au point la stratégie de « *DNA Shuffling* », nous avons décidé de faire un essai de *Shuffling* sur la banque d'*αRep* de première génération qui a été à son tour filtrée.

3. Essais de « Shuffling » de la banque de première génération

La banque d'*αRep* de première génération ne comporte que 30% de clones codants et 60%, seulement, des motifs insérés ont une séquence conforme. Comme présenté au préalable, notre démarche globale pour l'amélioration de la qualité de nos banques d'*αRep* est d'éliminer les clones non-codants par *phage display* et par la suite récupérer leurs motifs qui seront mélangés dans le but de recréer de la diversité.

Cette stratégie a été appliquée à la banque de première génération, dans le but d'effectuer les mises au point nécessaire avant de passer à la construction de la banque de deuxième génération. Ainsi la banque de première génération a été filtrée puis « *shufflée* ». Les mises au point réalisées et le résultat de cette stratégie seront exposés dans la partie qui suit.

3.1. Filtration de la banque de première génération

Comme présenté dans le Chapitre I, la banque de première génération a été construite dans un plasmide d'expression / *display* qui code une étiquette *Strep-tag* à l'extrémité N-terminale de l'*αRep* et une deuxième étiquette *His-tag* à l'extrémité C-terminale. Une fois,

l'*αRep* exposée à la surface du phage, le *Strep-tag*, situé à son extrémité N-terminale sera accessible et peut ainsi interagir avec la *Strep-Tactin* ce qui peut être un moyen efficace pour séparer les phages exposant des clones codants de ceux qui exposent une protéine tronquée. Ces constructions comportent également un *His-tag* qui aurait pu être utile pour réaliser cette même opération. Le choix du *Strep-tag* est basé essentiellement sur son emplacement. Il est en effet plus accessible à l'interaction que le *His-tag* qui se trouve entre les 2 parties de la protéine de fusion (Strep-tag_α-Rep_His-tag_pIII) et de ce fait probablement moins accessible.

Un essai de filtration de la banque de première génération a été réalisé dans l'équipe, par Agathe Urvoas, en exposant la banque sur phages qui sont mis ensuite mis en contact avec la *Strep-Tactin*. Les phages spécifiques sont récupérés ce qui va représenter la ***banque filtrée***.

Afin de vérifier l'efficacité du processus de filtration, nous avons eu recours à une caractérisation des variants de la banque filtrée. Cette caractérisation est réalisée par l'étude des séquences de clones issus de la banque filtrée et par un test d'expression soluble.

Suite à l'analyse des séquences de clones de cette banque, 12 clones sur 16 séquencés (69%) ont une séquence codante. L'expression protéique en milieu liquide a été vérifiée pour ces 12 clones codants. La filtration grâce au *Strep-tag* en utilisant des phages dialysés nous a permis alors d'augmenter la fraction des clones codant de 30% dans la banque de première génération à 69% dans la banque filtrée. Nous avons alors cherché à vérifier si la filtration influe sur la distribution des tailles des protéines codées dans la banque. Nous avons alors comparé la taille des protéines de la banque de première génération naïve et celle filtrée.

Pour cela, les phagemides de la banque naïve et ceux de la banque filtrée ont été digérés NdeI-HindIII. Ces deux sites sont localisés de part et d'autre du *Ncap* et du *Ccap* ce qui permet de libéré le gène codant l'*αRep*. Les produits de cette digestion ont été par la suite séparés sur un gel d'agarose 1%.

La distribution de la taille des protéines de la banque filtrée, observée sur le gel, est nettement différente de celle de la banque naïve. En effet, les protéines à grand nombre de motifs insérés ont été éliminées et il y a un net enrichissement en variants comportant 1, 2 ou 3 motifs insérés. Ceci confirme l'hypothèse émise : plus une protéine est longue plus la probabilité d'insérer des motifs non-codants est élevée.

Cet essai nous a permis de conclure que la filtration de la banque d'*αRep* par le biais de l'une des étiquettes en particulier celle du coté N-terminale est tout à fait pertinent pour enrichir en clones codants. Toutefois, il faut noter qu'il existe toujours des clones non-codants. Ceci peut être dut à la faible affinité du *Strep-tag* à la *Strep-Tactin* d'une part et à l'hétérogénéité de séquence du tag qui a été révélée par spectroscopie de masse (Chapitre I). Dans la nouvelle banque qui sera construite, la filtration sera réalisée en se basant sur une interaction décrite comme plus affine : *Flag-tag* avec un anticorps anti-*Flag-tag* ce qui laisse espérer une efficacité de filtration améliorée.

Cette banque filtrée voit sa diversité nettement restreinte par rapport à la diversité de la banque naïve. Il est alors important de mettre au point la stratégie de recréation de la diversité par « *Shuffling* ».

3.2. *Shuffling* de la banque de première génération filtrée

Une banque filtrée voit sa diversité diminuée considérablement en particulier au niveau de la taille de protéines qui comprennent 0 à 3 motifs seulement. Notre stratégie consiste à extraire les motifs codants des clones filtrés pour ensuite les mélanger (*Shuffling*) afin de recréer de la diversité. Ainsi nous avons procédé à un essai de *Shuffling* sur la banque de première génération filtrée.

Comme première approche, nous avons digéré les plasmides avec l'enzyme de restriction BsmBI pour libérer les séquences codant chaque motif. Nous avons simultanément digéré avec une autre enzyme qui coupe dans 3 sites asymétriques situés ailleurs dans le plasmide (site BsaI). Ces sites sont illustrés sur la carte du phagemide du clone αRep-n4-a (Fig. 11).

La coupure avec le site BsmBI est effectuée pour libérer les motifs insérés tandis que la coupure par les sites secondaires (BsaI) vise à digérer le plasmide en fragments évitant ainsi sa refermeture sur lui-même lors de la ligation. La recircularisation du plasmide sans insertion de modules serait en effet très fortement favorisée si celui n'était coupé que par les sites inter-modules (BsmB1) puisque la refermeture serait alors une réaction intramoléculaire. Un traitement par BsmB1 seul suivi d'une ligation conduirait majoritairement à exciser les modules de la banque plutôt qu'à les recombiner. Le traitement par BsaI impose un caractère intermoléculaire à la reconstitution du plasmide circulaire et favorise de ce fait l'incorporation de modules avant l'évènement de refermeture du vecteur. Le caractère asymétrique des

séquences des extrémités cohésives issues des restrictions BsaI et BsmBI est ici essentiel : l'utilisation d'un site palindromique à la place des sites BsaI et BsmBI conduirait à un procédé de *shuffling* inopérant ; les ligations multi-fragments avec sites symétriques étant très peu efficaces de par la multiplicité des produits de ligation possibles.

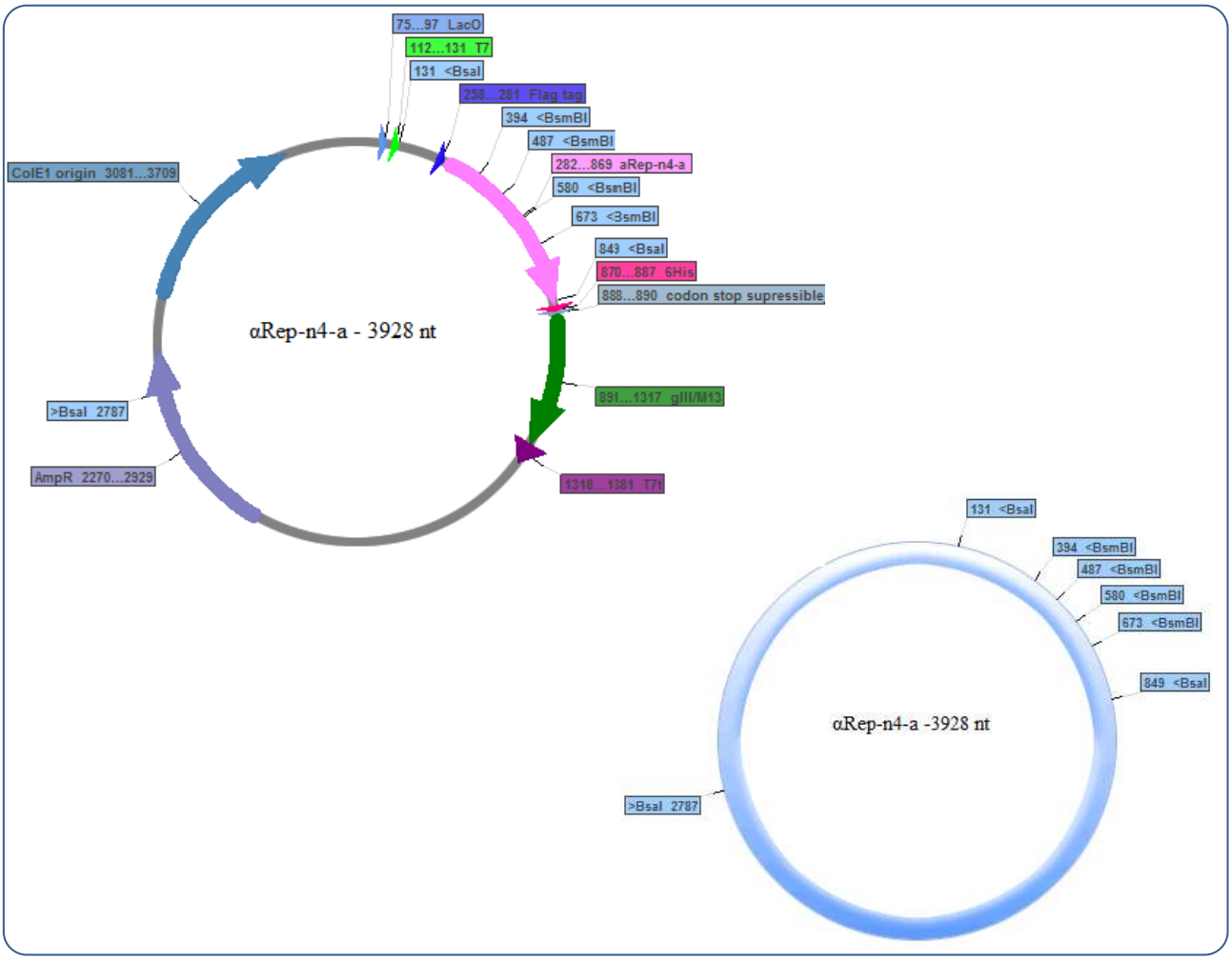


Fig. 11: Carte du phagemide du clone αRep-n4-a et les sites de restriction utilisés pour le « *Shuffling* » de la banque d'*αRep* de première génération filtrée. Le site BsmBI est utilisé pour libérer les motifs insérés. BsaI a 3 sites de restriction dans le plasmide ayant pour rôle le ralentissement de la fermeture du plasmide sur lui-même et l'insertion du plus grand nombre de motifs possible.

5 µg d'ADN de la banque filtrée ont été incubés avec 10 U de BsmBI et 10 U de BsaI (NEB) dans le tampon NEB3 pendant 1h30' à 50°C. Le produit de digestion est purifié par le kit NucleoSpin® ExtractII de chez Macherey-Nagel. Les fragments d'ADN sont élués dans 20 µL de tampon d'élution puis ils sont ligués dans 30 µL de volume total avec 800 U de T4 DNA ligase et 3 µL de tampon de ligase o.n. à 16°C. Le produit de ligation est purifié et élué dans 50 µL de tampon d'élution dilué 10 fois. Il a été ensuite utilisé pour transformer des XL1Blue MRF' par électroporation. 10 électroporations ont été réalisées : 100 µL de bactéries électrocompétentes sont mélangées avec 5 µL d'ADN, elles ont été soumises à un choc électrique puis resuspendues dans

*900 µL de milieu Soc et incubées 1h à 37°C. Les bactéries transformées ont été étalées sur milieu solide 2YTagar+Amp+Glu et 10 µL ont été prélevés pour compter le nombre de bactéries transformées. Ainsi nous avons obtenus une mini-banque de $4.2*10^7$ clones indépendants : banque d'**αRep de première génération filtrée shufflée**.*

Les bactéries de la banque ont été récupérées de la boîte de milieu solide puis conservées dans du milieu 2YT+Glu1%+20% Glycérol à -80°C. Un échantillon de cette banque filtrée shufflée a été utilisé pour extraire les phagemides représentant cette banque. Les phagemides de la banque naïve, la banque filtrée et celle shufflée ont été digérés NdeI-HindIII (Fig. 12) dans le but de visualiser et comparer la distribution de la taille des gènes codants les protéines des différentes banques.

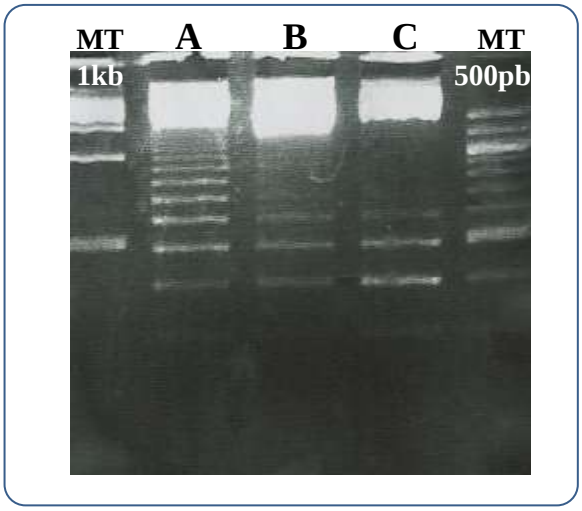


Fig. 12 : Electrophorèse comparant la distribution du nombre de motifs insérés par protéine dans la banque naïve (A), la banque filtrée (B) et la banque s*hufflée* (C).

La distribution de la taille des séquences des protéines issues des différentes banques, visualisée sur le gel précédent, montre clairement que la filtration diminue significativement le nombre de protéines ayant insérées 4 motifs ou plus. Le « *Shuffling* » bien qu'il permette d'introduire plus de diversité de séquence dans la banque filtrée, reste avec cette procédure, insuffisant pour augmenter la proportion de séquences longues.

A ce niveau, nous nous sommes proposés d'ajouter des inserts purifiés lors de la ligation des phagemides dans le but d'augmenter la diversité de la taille des protéines dans la banque.

Ainsi, on a procédé à 2 digestions BsaI-BsmBI de la même quantité de phagemides de la banque filtrée dans 2 tubes différents : une 1ére digestion est juste purifiée alors que l'autre est

séparée sur gel agarose 2% (Fig. 13) dans le but de purifier la bande de 100 pb correspondant aux inserts libérés.

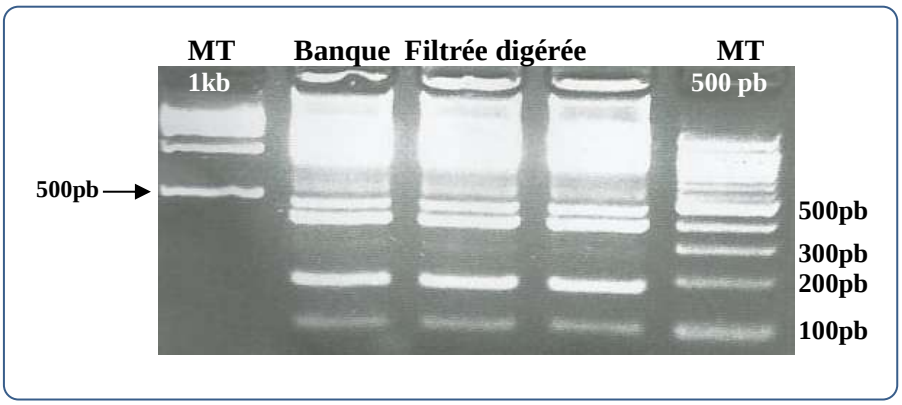


Fig. 13 : Séparation du produit de digestion BsmBI-BsaI des phagemides de la banque filtrée. La bande de 100pb correspond aux motifs insérés qui ont été purifiés du gel puis ajoutés à la ligation.

La ligation a été réalisée avec le produit de la première digestion et les motifs purifiés du gel. Le produit de ligation a été purifié, utilisé pour électroporer des XL1Blue MRF' et ainsi obtenir la mini-banque ***d'αRep de première génération filtrée shufflée2****. Une digestion NdeI-HindIII a été réalisée dans le but de comparer la distribution de la taille des protéines dans les 4 types de banques (Fig. 14).*

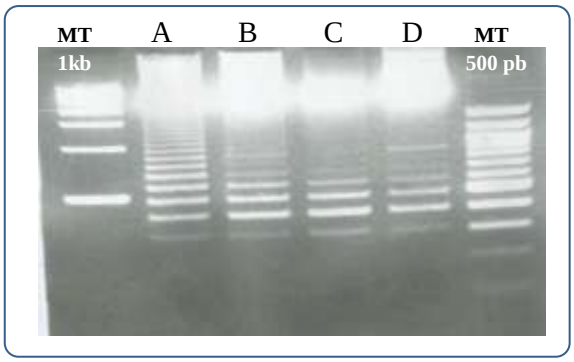


Fig. 14: Electrophorèse montrant la distribution du nombre de motifs insérés dans la banque naïve (A), la banque filtrée (B), la banque filtrée *shufflée* (C) et la banque filtrée shufflée2.

Ce profil montre que l'ajout d'un excès de motifs lors de la ligation permettait d'obtenir des séquences de taille plus importante et aussi d'enrichir en protéines à 4 et 5 inserts et atteindre des protéines à 7 motifs insérés. Ces constatations illustrées sur le profil de digestion des plasmides, sont confirmées par l'extraction des plasmides de 19 clones pris au hasard dans la mini-banque d'*αRep* de première génération filtrée et shufflée2 : 10% des clones à 0 motif (2 clones), 26% des clones à 1 motif (5 clones), 21% à 2 motifs (4 clones), 42% à 3 motifs (8 clones) et on a même des protéines à 7 motifs insérés.

Ces résultats montrent que la filtration permet d'augmenter la fraction des protéines à séquence codante mais elle réduit nettement la diversité utile de la banque en réduisant la gamme de taille des protéines vers celles ayant un ou deux motifs insérés. Cet inconvénient a été par la suite compensé en recréant de la diversité par mélange des motifs et/ou en introduisant un excès de motifs ce qui permettrait d'augmenter la taille des protéines.

Ces essais, nous ont permis de mettre au point une stratégie générale pour la construction de la banque d'*αRep* de deuxième génération. L'objectif est que cette nouvelle banque soit optimale à deux niveaux : diversité des positions hypervariables (diversité qui mime la diversité naturelle) et fraction des variants codants (grand nombre de variants codants et on vise d'atteindre un 100% de codants). Si les méthodes de filtration et *shuffling* décrites ci-dessus permettent de l'améliorer, il est néanmoins essentiel de concevoir et construire une banque primaire qui soit, avant filtration et s*huffling*, la meilleure possible. Nous avons pour cela introduits des améliorations dans la procédure de construction de la banque à deux niveaux :

Tout d'abord, les gènes synthétiques codant pour les cercles, qui seront amplifiés par RCA, seront conçus de sorte à coder la diversité naturelle pour les positions 18, 19, 22, 23, 26 et 30 plutôt qu'une diversité biaisée un peu arbitrairement. Ensuite, à la différence avec la banque de première génération, les cercles amplifiés seront intégralement doubles brins dans le but de diminuer les risques d'hybridation non spécifique ou de structure locales et par la suite de minimiser les erreurs d'assemblage ou d'amplification des séquences recherchées. Ces améliorations sont détaillées dans les pages qui suivent.

Conclusion

Une banque de deuxième génération 2.0 a été construite par Agathe Urvoas en utilisant le vecteur accepteur avec *pLAcT7prom* et le *Flag-tag* et des cercles de motifs obtenus par hybridation d'oligonucléotides doubles brins (Résultats non publiés). Cette banque a été conçue de sorte que la diversité dans les positions hypervariables mime la diversité naturelle. Elle a été aussi caractérisée : nombre de variants codants et la distribution des tailles des *αRep*. Cette banque ne contient que $7.5*10^7$ mutants dont la taille varie entre 0 et 3 motifs insérés avec un grand nombre de variants ne contenant que des *Ncap* et des *Ccap*. Cette banque nous a permis de vérifier que l'utilisation d'oligonucléotides doubles brins lors de la préparation des cercles permettait en effet d'augmenter la proportion de clones codants de la banque.

Nous avons, alors, décidé de construire une banque de deuxième génération 2.1 selon la stratégie générale suivante :

- **Une *banque primaire* serait construite en utilisant les mêmes cercles de la banque de deuxième génération2.0.**
- **Les clones codants seront filtrés par phage display : *banque primaire filtrée.***
- **Les motifs de la banque primaire filtrée seront extraits et mélangés : *la banque filtrée shufflée = banque de deuxième génération 2.1.***

C. Construction de la banque d'αRep de deuxième génération 2.1

Introduction :

Toutes les optimisations réalisées au préalable nous ont permis de fixer la stratégie générale pour la construction de la nouvelle banque d'*αRep*. La procédure de construction implique plusieurs étapes successives: il faut tout d'abord obtenir une banque primaire qui sera par la suite « filtrée » pour éliminer les clones ayant des mutations non-sens. La filtration ne conservant que la diversité utile de la banque primaire, diminuera sensiblement le nombre de clones et conduira à une diversité limitée. Il sera alors utile de recréer de la diversité par «*Shuffling* » des motifs préalablement sélectionnés pour être corrects.

Ces étapes seront décrites dans cette partie de la thèse puis nous décrirons les caractéristiques de la banque que nous avons effectivement obtenue en suivant cette stratégie. Nous décrirons enfin l'utilisation de cette banque pour identifier des *αRep* interagissant avec des protéines cibles préalablement choisies.

1. Construction moléculaire de la banque primaire

La procédure de construction de la bibliothèque est similaire à la procédure utilisée pour construire la bibliothèque décrite dans le chapitre I mais comporte quelques adaptations qui seront détaillées ici. La construction moléculaire nécessite plusieurs étapes : la conception puis la préparation du vecteur accepteur (Chapitre II-A), la synthèse des motifs à insérer, l'intégration de ces motifs dans le vecteur, puis l'obtention d'une population de bactéries transformées par électroporation.

1.1. Conception et préparation des cercles d'ADN codant les motifs *αRep*

Les fragments d'ADN codant les motifs répétés sont produits par amplification circulaire isotherme (RCA). Cette amplification est basée sur l'amplification préférentielle de fragments d'ADN circulaires. Les « cercles » sont constitués par hybridation d'une série d'oligonucléotides. Puis une fois circularisés et ligaturés, ils peuvent servir efficacement de matrice pour l'amplification par RCA. Les assemblages incorrects ou incomplets, qui ne conduisent pas à un cercle ne sont pas amplifiés. Chaque cercle codera pour un motif et comportera donc les régions constantes du motif et une séquence partiellement aléatoire pour les codons dégénérés. Deux types d'oligonucléotides nucléotides ont été conçus pour couvrir toute la séquence du motif (Fig. 1) :

- des oligonucléotides correspondant aux régions constantes notés C : au nombre de deux représentés sur la séquence du motif en noir : ***C-1-for*** et ***C-1-rev***.

- des séries d'oligonucléotides correspondant aux régions variables notés V : 3 séries ont été conçues selon la position codée :
  - ✓ 25 oligonucléotides ***Va*** et leur complémentaires, représentés en rouge sur la séquence du consensus, sont conçus de sorte à rendre variables les positions 18 et 19.
  - ✓ 24 oligonucléotides ***Vb*** et leur complémentaires, représentés en vert sur la séquence du consensus, sont conçus de sorte à rendre variables les positions 22 et 23.
  - ✓ 16 Oligonucléotides ***Vc*** et leur complémentaires, représentés en bleu sur la séquence du consensus, sont conçus de sorte à rendre variables les positions 26 et 30.

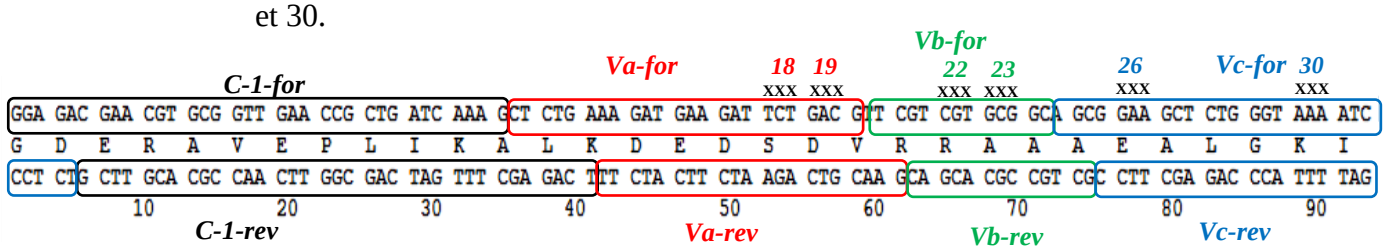


Fig. 1 : Séquence d'un motif consensus représentant l'emplacement des oligonucléotides conçus pour coder la totalité du motif. Les oligonucléotides *Va* et *Va-rev* permettent de rendre les positions 18 et 19 variables par des mutations dans les nucléotides XXX. De même, les oligonucléotides *Vb* et *Vc* et leurs complémentaires mutent respectivement les positions (22, 23) et les positions (26, 30). *C-1-for* et *C-1-rev* sont les deux oligonucléotides conçus pour coder la région conservée du motif.

Plusieurs modifications ont été apportées à la procédure utilisée lors la construction de la banque d'αRep de première génération.

- Tout d'abord, les cercles obtenus par hybridation et ligation sont des cercles doubles brins, et il ne subsiste plus de régions simples brins lors de l'amplification, afin de minimiser des sites d'hybridation erronés et/ou la formation d'éventuelles structures en boucle dans les régions simples brins pouvant entrainer des erreurs de polymérisation.

- Chaque position hypervariable (18, 19, 22, 23 26, et 30) est codée par un ensemble de codons dégénérés. Il est en effet essentiel de créer une diversité importante mais sans pour autant parvenir à une distribution d'acides aminés qui soit complétement aléatoire. Certains acides aminés peuvent être déstabilisants dans un environnement particulier. On conçoit bien, par exemple, que l'incorporation d'un résidu proline au milieu d'une hélice α va sévèrement déstabiliser la protéine dans lesquelles ce résidu se retrouvera. Il est donc essentiel d'éliminer les résidus prolines au sein des hélices. En revanche, ce même résidu proline peut être stabilisant s'il est inséré à une extrémité d'hélice où il peut participer à une structure de type

« *capping-box* ». Plus généralement, cela suppose de définir quelles sont les chaines latérales qui sont effectivement tolérées pour chaque position hypervariable, et non globalement pour toutes les positions.

Dans ce but, la distribution des fréquences des 20 types d'acides aminés à chaque position hypervariable a été calculée en utilisant comme base statistique une collection de 1710 motifs *αRep* naturels. Ces motifs ont été extraits de 500 protéines non redondantes détectées par similitude à un déca-consensus *αRep*, dans lequel les positions aléatoires ne sont pas spécifiées. La distribution des acides aminés montre clairement que ces positions si elles sont hypervariables, elles ne sont pas aléatoires et les biais observés différent à chaque position. Par exemple la proline est le résidu le plus souvent observé en position 18, à l'extrémité N de l'hélice 2, alors qu'il est presque totalement absent aux positions 22, 23 et 26 qui se trouvent incluses dans l'hélice 2. D'autres biais sont également très nets : Par exemple, la position 23 est très fréquemment occupée par des chaînes latérales de petites tailles, ce qui s'explique par son orientation vers l'hélice α adjacente et l'encombrement stérique qui en résulte. D'autres biais sont plus difficiles à expliquer comme les très faibles fréquences de la méthionine, de la cystéine et de l'histidine. Il est concevable que ces acides aminés soient naturellement contre-sélectionnés du fait d'une instabilité chimique liée à leur réactivité propre, particulièrement dans des microorganismes ayant un caractère thermophile, ce qui semble fréquent parmi les organismes sources de ces protéines.

Nous avons donc cherché à reconstituer une diversité qui s'approche de la diversité naturelle position par position. Ceci n'est pas réalisable en utilisant les schémas de dégénérescence reposant sur les codons NNK ou NNS classiquement utilisés. Le codage de telles distributions serait possible en utilisant des oligonucléotides synthétisés à partir de précurseurs de trinucléotides (plutôt que de mononucléotides). Ces technologies de synthèse, bien que décrites (Virnekäs et *al.*, 1994), ne sont pas facilement disponibles et demeurent nettement plus onéreuses que les synthèses classiques.

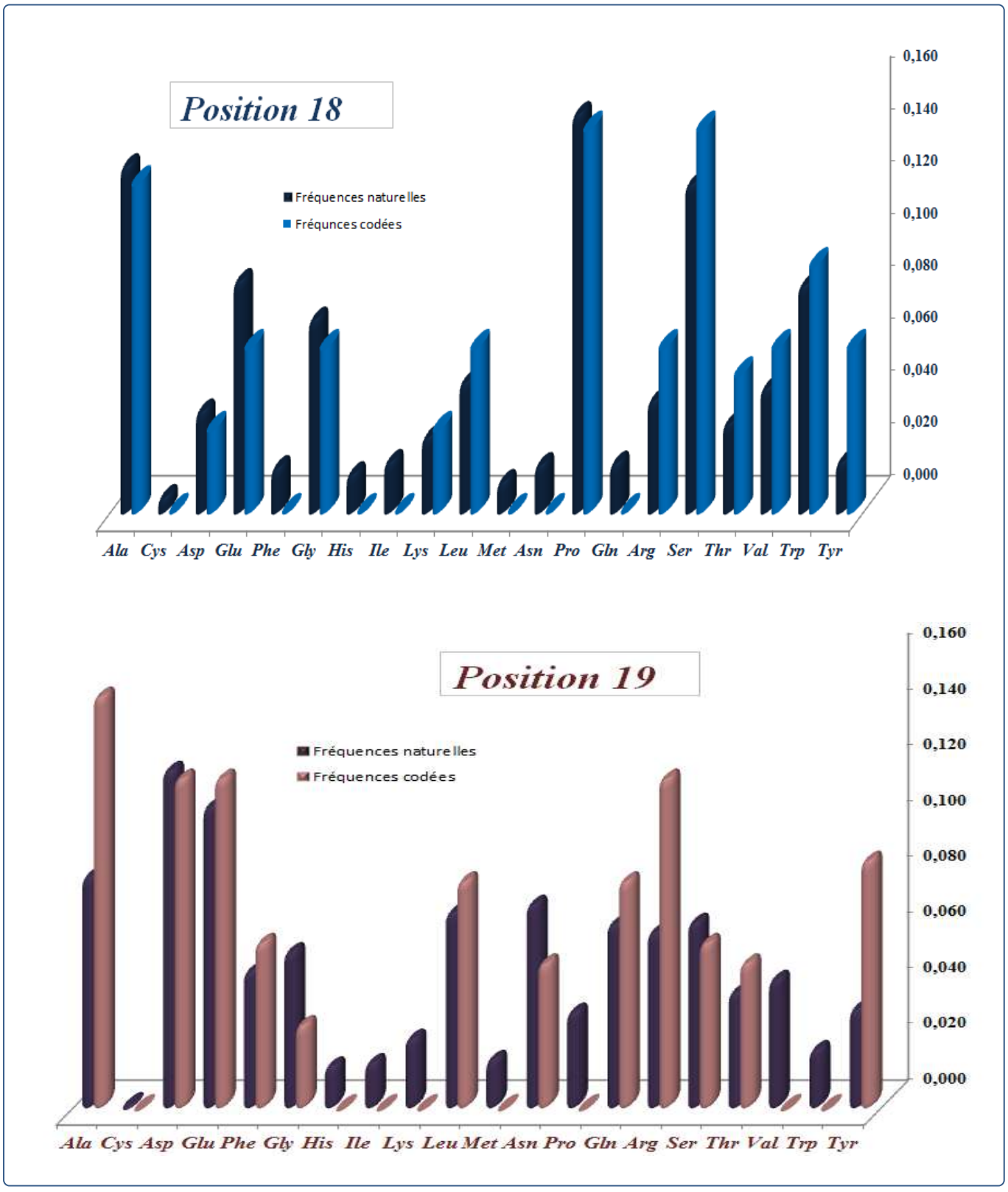
Nous avons donc cherché à coder chaque distribution par un mélange d'oligonucléotides, ou chacun d'entre eux comporte un codon partiellement dégénéré. Des distributions de codons partiellement dégénérés peuvent permettre de s'approcher efficacement des distributions cibles. La nature des codons partiellement dégénérés et leur proportion ont été définies, non par un calcul d'un codage optimal, mais simplement par essais erreurs successifs en simulant à partir du code génétique les distributions qui

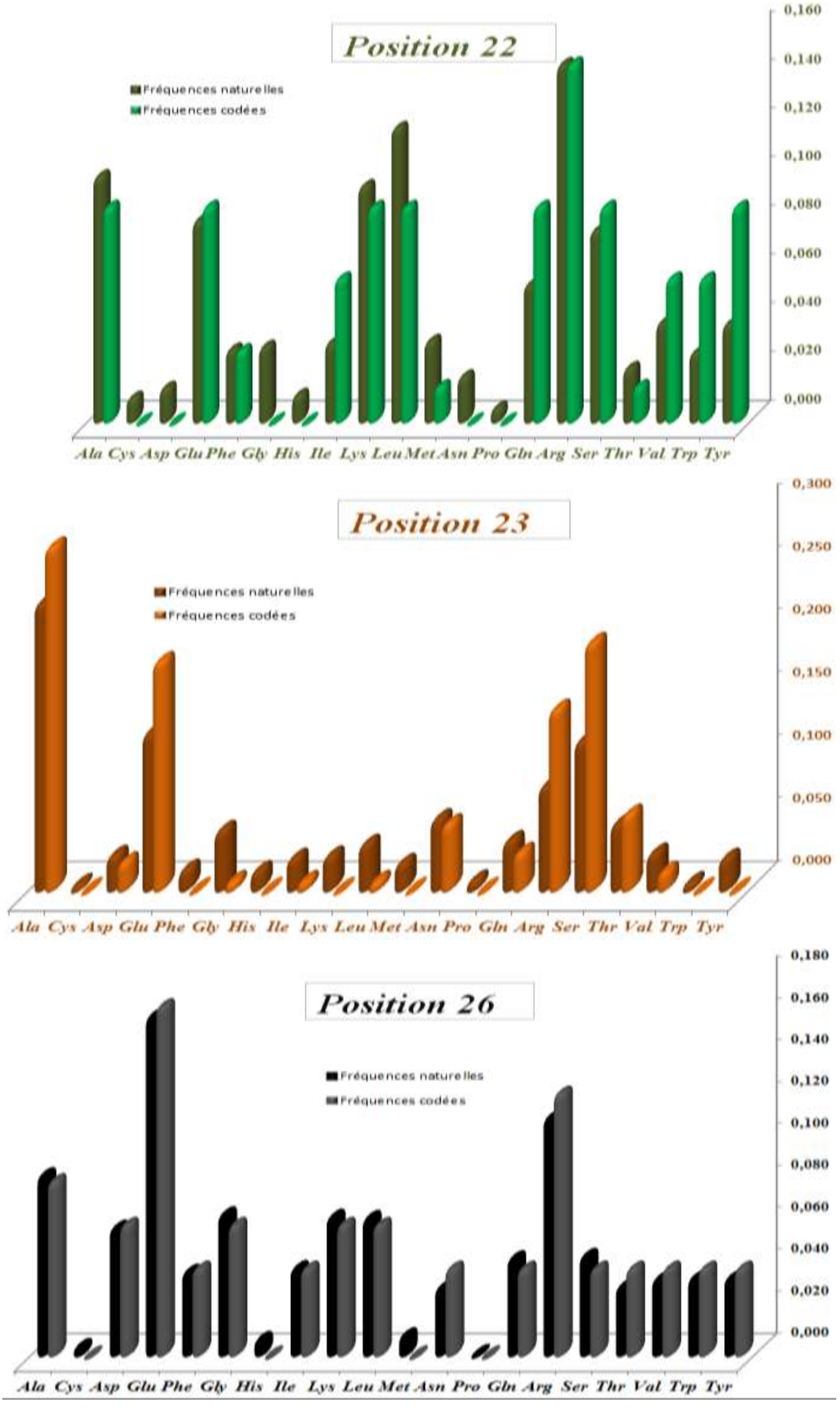
résulteraient de chaque codage envisagé. Dans le schéma retenu, 13 acides aminés ont été codés pour la position 18, 12 pour la position 19, 14 à la position 22, 12 pour la pistions 23, 16 à la position 26 et seulement 3 pour la position 30.

Nous avons sur-représenté les acides aminés aromatiques relativement à leur fréquence naturelle aux positions 18, 19, 22, 26, parce que ces résidus contribuent de façon fréquente aux interfaces protéine/protéine. Cela n'a pas été fait à la position 23 pour des raisons stériques ni à la position 30 dont la séquence nucléotidique est contrainte par la présence du site de restriction utilisé. Le tableau suivant résume les codons, ou paires de codons, dégénérés utilisés et les acides aminés pour lesquels ils codent :

Codon 18	Acides aminés 18	Codon 19	Acides aminés 19	Codon 22	Acides aminés 22	Codon 23	Acides aminés 23	Codon 26	Acides aminés 26	Codon 30	Acides aminés 30
hcc	P/S/T	gac	D	cgt	R	gct	A	Gaa	E	vaa	E/K/Q
kac	D/Y	gma	A/E	cgt	R	gaa	E	Cgt	R	vaa	E/K/Q
bcg	A/P/S	kcg	A/S	cgt	R	aac	N	Gcg	A	vaa	E/K/Q
bcg	A/P/S	gma	A/E	cgt	R	ryt	I/T/V/ A	Ggc	G	vaa	E/K/Q
bcg	A/P/S	cwg	L/Q	cgt	R	skt	G/R/L/ V	aaa	K	vaa	E/K/Q
bcg	A/P/S	amc	N/T	raa	K/E	gma	A/E	ctg	L	vaa	E/K/Q
bcg	A/P/S	twc	F/Y	raa	K/E	rmc	N/T/D/ A	gat	D	vaa	E/K/Q
bcg	A/P/S	cgc	R	kct	S/A	gma	A/E	agc	S	vaa	E/K/Q
tgg	W	cwg	L/Q	kct	S/A	sag	Q/E	cag	Q	vaa	E/K/Q
tgg	W	cgc	L	kct	S/A	mgc	R/S	att	I	vaa	E/K/Q
gwa	E/V	gma	A/E	cwg	L/Q	gma	A/E	ttc	F	vaa	E/K/Q
gwa	E/V	amc	N/T	cwg	L/Q	mgc	R/S	tgg	W	vaa	E/K/Q
gwa	E/V	twc	F/Y	cwg	L/Q	wct	S/T	tat	Y	vaa	E/K/Q
gwa	E/V	cgc	R	tac	Y	gcg	A	gtg	V	vaa	E/K/Q
sgt	G/R	cwg	L/Q	tac	Y	mgc	R/S	acc	T	vaa	E/K/Q
sgt	G/R	tac	Y	tac	Y	ggt	G	aac	N	vaa	E/K/Q
sgt	G/R	gma	A/E	ttc	F	tcg	S				
sgt	G/R	twc	F/Y	tgg	W	cgt	R				
ctg	L	ggt	G	atc	I	gma	A/E				
tac	Y	tct	S	atc	I	mgc	R/S				
ctg	L	gma	A/E	tgg	W	kct	A/S				
ama	K/T	gma	A/E	ayg	M/T	kct	A/S				
ama	K/T	amc	N/T	gtt	V	gcg	A				
ama	K/T	cwg	LQ	gtt	V	mgc	R/S				
ccg	R	tgg	W								
ncg	A/P/S/ T	gtt	V								

Nous représentons également sur les histogrammes suivants la comparaison entre la diversité naturelle et celle codée par les codons dégénérés retenus, et cela pour chacune des positions hypervariables (18, 19, 22, 23, 26 et 30) (Fig. 2).




Position 22
Fréquences naturelles
Fréquences codées
0,160
0,140
0,120
0,100
0,080
0,060
0,040
0,020
0,000
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr
Position 23
Fréquences naturelles
Fréquences codées
0,300
0,250
0,200
0,150
0,100
0,050
0,000
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr
Position 26
Fréquences naturelles
Fréquences codées
0,180
0,160
0,140
0,120
0,100
0,080
0,060
0,040
0,020
0,000
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr

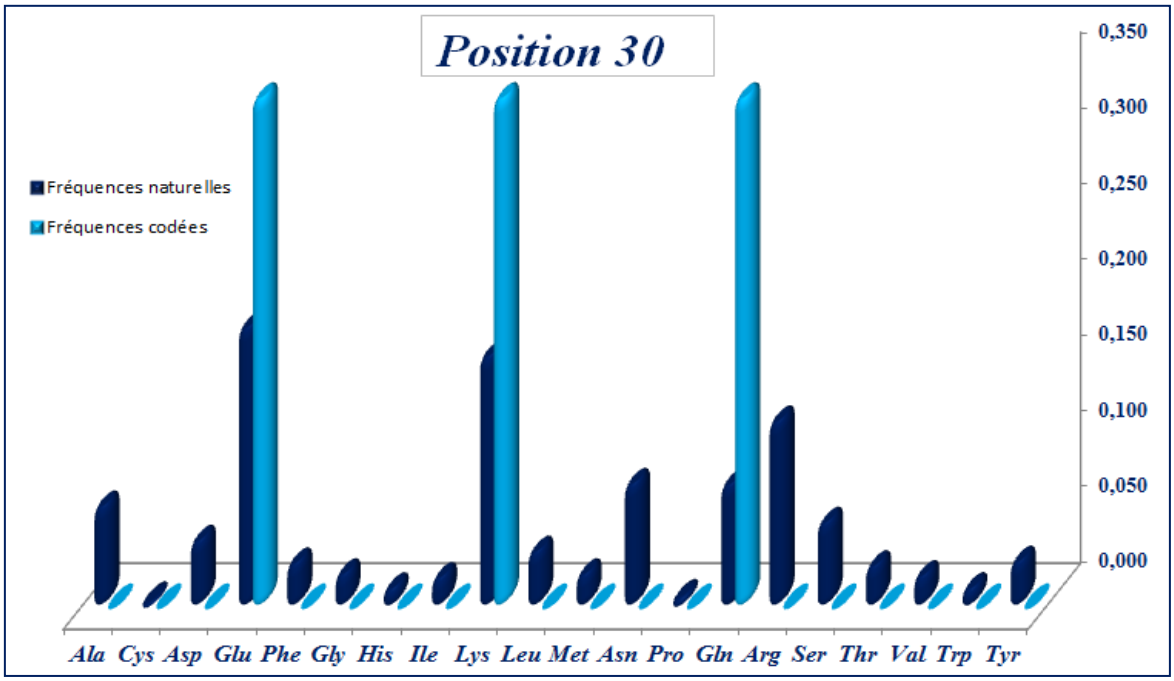


Fig. 2 : Comparaison entre la distribution naturelle et celle codée par les oligonucléotides *Va*, *Vb* et *Vc* des acides aminés aux positions hypervariables.

Fréquence des dipeptides contigus

Les positions 18 et 19 d'une part, et 22 et 23 d'autre part, sont contiguës et ces paires de codons se trouvent par conséquent chacune codée sur un seul oligonucléotide. Cela introduit une combinatoire entre les codons retenus à la position 18 et ceux à la position 19 (ou entre 22 et 23) qui va conduire à synthétiser autant d'oligonucléotides que de paires de codons dégénérés. Nous n'avons échantillonné que partiellement cette combinatoire en privilégiant les paires les plus fréquentes. Par exemple, 10 codons dégénérés ont été utilisés pour la position 18 et 10 pour la position 19. Cela génère théoriquement 100 couples (18,19) et donc 100 oligonucléotides (et autant sur la chaîne complémentaire). En pratique, pour la paire (18,19), seuls 25 oligonucléotides ***Va*** (et 25 ***Va-rev***) ont été synthétisés. Cet ensemble permet de coder pour 87 dipeptides différents, soit un peu plus de la moitié des 156 dipeptides théoriquement possibles avec les acides aminés retenus à chaque position. Nous avons comparé la distribution des fréquences des dipeptides 18-19 résultant du schéma de codage retenu avec la distribution des dipeptides observés naturellement : 65 des 87 dipeptides codés dans la bibliothèque se retrouvent aussi parmi les 100 dipeptides les plus fréquemment rencontrés naturellement à ces positions. Réciproquement, 64 des 100 dipeptides naturellement les plus fréquents à ces positions sont représentés dans la bibliothèque. Les dipeptides codés dans la bibliothèque correspondent donc principalement aux dipeptides qui sont naturellement les plus fréquents. Cela résulte de la conception de la bibliothèque qui

élimine les acides aminés rares et représente principalement les acides aminés les plus fréquents. Or, dans la collection de modules naturels la fréquence naturelle d'un dipeptide observé est souvent proche du produit des fréquences de chaque acide amine. En d'autres termes, les effets de covariance de résidus voisins semblent généralement limités. Quelques anomalies ont été observées : par exemple, le dipeptide 18-19 le plus fréquent (WQ) est observé naturellement beaucoup plus fréquemment (F (W, Q) = 31/1000) qu'attendu par le produit des fréquences de ces deux résidus (F(W) x F(Q) <6/1000). Cela suggère, dans ce cas précis, une covariance de ces deux types de résidus. Cette anomalie a été mimée par un oligonucléotide qui inclut le codage de cette paire particulière.

De la même façon, le dipeptide 22-23 est codé par une série de 24 oligonucléotides qui codent 59 dipeptides différents. La paire (22,23) est naturellement moins diverse que la paire (18,19) du fait du biais en faveur des petits résidus à la position 23. De ce fait seul 89 dipeptides différents sont représentés dans la collection de modules naturels. La distribution de codons dégénérés, que nous avons retenus pour les positions 22 et 23, code pour 39 des 50 dipeptides naturellement les plus fréquents, et seuls 6 des dipeptides codés par les oligonucléotides ***Vb*** ne se rencontrent pas parmi les 50 dipeptides les plus fréquents

Le nombre de modules réellement échantillonnés en tenant compte de diversité des dipeptides 18-19, des dipeptides 22- 23 ainsi que des positions 26 et 30 est donc de 2,46 $x10^5$ modules différents (87 x 59 x 16 x 3). Cette combinatoire est de taille relativement réduite par rapport à la taille expérimentale des banques accessibles (10^8) et peut donc être échantillonnée exhaustivement. Mais une protéine comportera plusieurs modules aléatoirement associés ce qui conduit à une diversité théorique bien supérieure. Par exemple un tel schéma peut conduire à $1,5*10^{16}$ protéines différentes comportant 3 modules, soit beaucoup plus qu'il n'est possible d'échantillonner expérimentalement.

Pratiquement, les oligonucléotides conçus ont été synthétisés chez MWG eurofins et purifiés par HPLC. Chaque couple d'oligonucléotides complémentaires (for et rev) est hybridé séparément. Les différentes cassettes doubles brins sont mélangées dans les proportions requises puis liguées à 16°C pour former le motif entier. La concentration relative des différents ***Va*** *ainsi que des différents* ***Vb*** *sont fixés de sorte à s'approcher des distributions recherchées. La concentration des cassettes cercles correspond à une concentration potentielle de cercles égale à 2 µM. La ligation va permettre alors l'obtention des cercles qui sont utilisés comme matrice à l'étape suivante d'amplification (RCA).*

1.2. Obtention des motifs αRep par amplification isotherme des cercles

Chaque cercle formé par l'hybridation des différents oligonucléotides code pour un motif *αRep*. Dans le but d'obtenir la quantité de motifs *αRep* nécessaires pour la construction de la banque avec la diversité en taille de protéines, il est primordial de disposer d'une quantité suffisante de fragments et donc d'amplifier ces cercles par RCA. Cette amplification isotherme (30°C) va permettre l'obtention d'une grande quantité d'ADN (µg) à partir de quelques picogrammes d'ADN matrice circulaire en un temps réduit (4h), avec une haute-fidélité de réplication. A l'issus de cette réaction, des concaténéres doubles brins et de haut poids moléculaires sont obtenus et ils correspondent à des répétitions des différentes matrices circulaires utilisées. Une ultime étape de digestion (BsmBI) va permettre de libérer les motifs de chaque motif en fragments individuels (93 bp chaque) qui seront par la suite hétéro-polymérisés pour constituer la bibliothèque.

*A la concentration de la solution de cercles doubles brin obtenus (2 µM), 1 µL de cette solution contient ainsi 1.2*10^{12} molécules, ce qui recouvre largement la diversité théorique calculée. 10 amplifications ont été réalisées, en parallèles avec le kit TempliPhi^{TM} de GE, comme suit :*

- *1 µL de cercles doubles brins ont été mélangés avec 10 µL du Sample buffer du kit puis ils ont été incubés 3' à 95°C. Ce tampon va permettre de reprendre l'échantillon d'ADN et plus important encore il contient les hexamères aléatoires qui vont servir d'amorces pour l'étape suivante d'amplification.*
- *L'hybridation des amorces du Sample buffer (hexamères aléatoires) a été réalisée par diminution de la température jusqu'à 4°C.*
- *On a ajouté par la suite 10µL du Reaction buffer et 0.4 µL d'Enzyme Mix. Le Reaction buffer est le tampon qui procure les sels, les deoxynucléotides et le pH convenables pour la synthèse. L'« Enzyme Mix » contient quant à lui la Φ29 polymérase et des hexamères aléatoires.*
- *L'amplification a été par la suite réalisée par incubation 4 h à 30°C.*

Si l'amplification est efficace, chaque cercle amplifié va former une structure homopolymérique de très grande taille. La réaction conduit donc à un mélange d'homopolymères. Ces homopolymères peuvent être coupés par le site BsmBI qui se trouve entre chaque motif pour former des oligomères de taille plus réduite analysables sur gel, jusqu'à des monomères linéaires lorsque la coupure est totale. Si l'assemblage des

oligonucléotides ou l'amplification est non-spécifique, la coupure devrait donner lieu à une distribution continue de fragments de taille quelconque, tandis que si l'assemblage et l'amplification sont spécifiques, on attend une distribution discrète de fragments dont la taille correspond à un nombre entier de motifs répétés (n*93 bp).

Dans le but de vérifier l'efficacité de l'amplification et obtenir des motifs individuels, une digestion BsmBI a été alors réalisée :

- *La Φ29 polymérase a été tout d'abord inactivée par chauffage 15 min à 65°C.*
- *On a ajouté par la suite 70 µL d'H_2O et 10 µL de tampon 50 mM Tris-HCl, 100 mM NaCl, 10 mM MgCl2, 1 mM Dithiothreitol à pH7 (tampon NEB 3) et 10 U de BsmBI dans chaque tube d'amplification.*
- *Le mélange réactionnel a été incubé à 55°C pendant une 1h30 min.*

Le produit d'amplification de chaque tube, digéré BsmBI, a été contrôlé par dépôt sur gel d'agarose 2 % : 2 µL ont été prélevés à 5 min d'incubation puis 2 µL ont été prélevés à la fin de la réaction (Fig. 3)

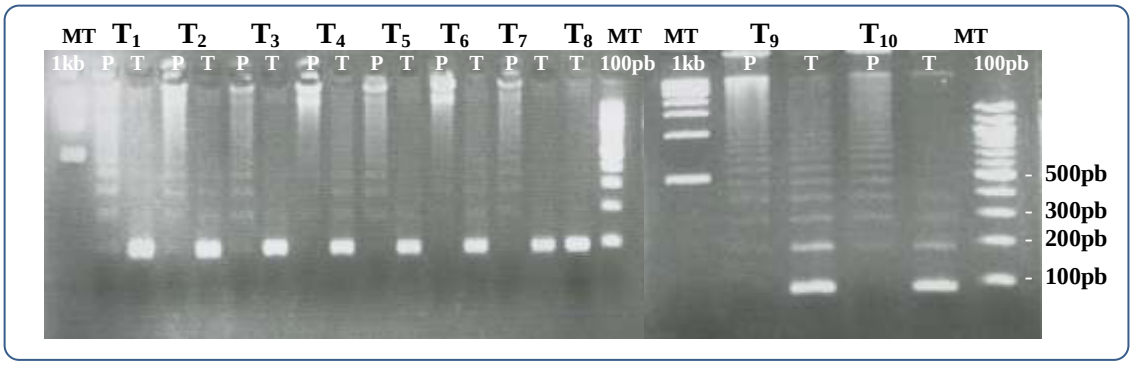


Fig. 3 : Profil de la digestion BsmBI du produit d'amplification pour les 10 tubes (de T_1 à T_{10}): (P) correspond à la digestion partielle (5min d'incubation à 55°C) et (T) correspond à la digestion totale.

Nous avons bien vérifié que les cercles ont été efficacement amplifiés. Puisque les polymères produits par RCA sont visibles dans la digestion partielle et ils ont été fragmentés en inserts individuels par digestion totale.

Ces fragments obtenus ont été purifiés à l'aide du kit ExtractII de Macherey Nagel puis ils ont été élués en 2 temps avec 15 µL de tampon d'élution. Nous avons obtenus une solution finale de 150 µL de motifs purifiés d'une concentration de 193.5 ng/µL ≈ 29µg de produit RCA. Ces motifs seront par la suite insérés dans le vecteur accepteur présenté dans la section (A). Ce dernier a été conçu pour permettre une surexpression dans la souche Rosetta Blue et BL21 d'E.coli et l'exposition sur les particules de phages M13.

1.3. Préparation du vecteur accepteur semi-biotinylé :

La procédure suivie pour l'insertion des motifs amplifiés est différente de la méthode classique de clonage utilisant deux sites de restriction. En effet, notre démarche a pour but de permettre la polymérisation des modules dans le vecteur. La polymérisation doit permettre l'incorporation de modules successifs ayant tous la bonne orientation, ce qui suppose d'utiliser un site de restriction de type 2s (BsmBI) non palindromique, dont les extrémités cohésives non symétriques ne peuvent s'apparier que dans une seule orientation. Les extrémités cohésives compatibles avec celle des modules doivent également se trouver sur le vecteur mais il est important d'éviter la fermeture du plasmide sur lui-même sans intégration d'inserts (Fig. 4).

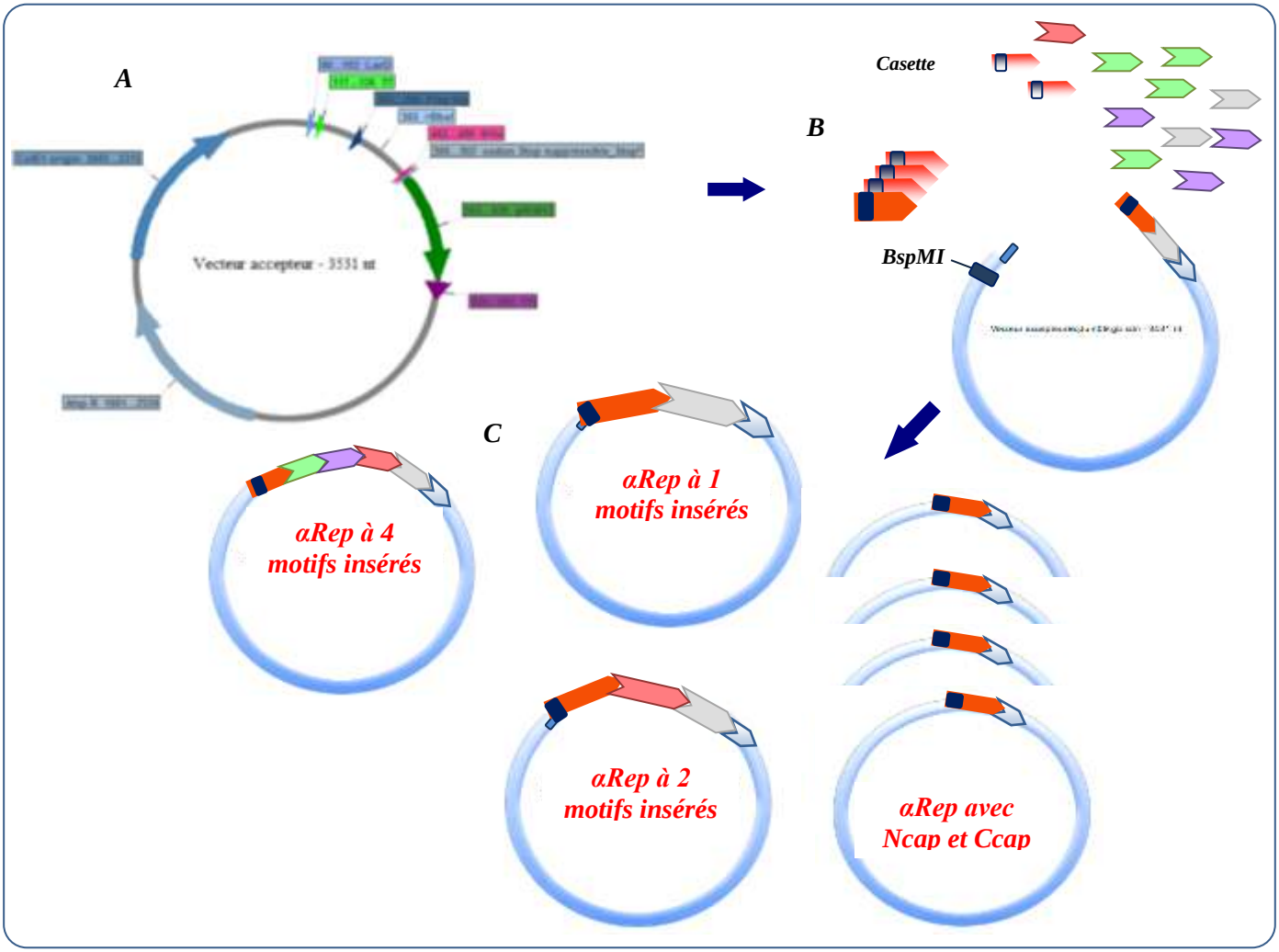


Fig. 4 : Exemple d'une procédure de ligation conventionnelle avec un site de restriction (BbsI) : A : le vecteur accepteur de la banque est tout d'abord digéré BbsI qui libère des bouts cohésifs du plasmide compatibles avec les bouts des motifs libérés par BsmBI. B : Ligation du plasmide digéré BbsI en présence d'un excès de motifs à insérer du motif N-cap. C : Le produit de ligation va contenir une majorité de plasmides qui va se refermer sans intégrer de motifs et une minorité qui va se liguer en ayant intégré 1 motifs ou plus.

Il est donc critique de ne libérer dans un premier temps qu'une seule des deux extrémités cohésives du vecteur, et de protéger la seconde extrémité de sorte qu'elle ne soit pas disponible pour une recircularisation, tant que la polymérisation des modules n'est pas réalisée. Cette libération successive des deux extrémités du vecteur peut être réalisée comme décrit dans l'article du chapitre I en clivant tout d'abord le vecteur par un enzyme qui génère l'une des extrémités cohésives nécessaire pour intégrer les modules, mais en éliminant la seconde extrémité cohésive du vecteur grâce à un site introduit pour cela (KpnI). Le vecteur n'ayant plus qu'une moitié de son site ne peut plus à cette étape se refermer sur lui-même et le module peut polymériser dans le vecteur. La refermeture du vecteur ne sera effectuée que dans un second temps en libérant, après polymérisation et insertion du cap, la seconde extrémité nécessaire pour la refermeture du vecteur.

Relativement à la procédure décrite dans l'article du chapitre I, une étape supplémentaire a été ajoutée. Le but de cette nouvelle étape est d'obtenir le vecteur utilisé pour la polymérisation sous une forme biotinylée. Cela permet de purifier très efficacement le vecteur ayant (ou non) incorporé les modules de l'excès de fragment ajouté lors de la polymérisation, et qui n'ont pas été incorporés. Cette étape n'est pas strictement indispensable mais permet de diminuer la quantité d'ADN inutile (les fragments en excès non-incorporés dans le vecteur) et de n'avoir lors de l'électroporation que les constructions utiles, ce qui permet d'augmenter les concentrations utilisées sans créer d'arc électrique.

Ainsi, cette étape permet l'amélioration de l'efficacité globale de la construction. La procédure générale comporte donc les étapes suivantes

- Coupure du site de protection (KpnI)
- Fixation de *linkers* biotinylés sur les extrémités libérées par KpnI
- Déprotection d'une des extrémités cohésives du vecteur (BbsI) coté *Ccap*
- Polymérisation des modules sur cette extrémité
- Addition du *Ncap* en excès
- Capture sur billes magnétiques streptavidine du vecteur biotinylé ayant incorporés les modules (et élimination des modules et *Ncap* en excès)
- Elution du vecteur capturé par restriction ce qui conduit simultanément à la « déprotection » de la seconde extrémité
- Refermeture par ligation intramoléculaire
- Electroporation

La procédure est détaillée dans la partie qui suit et elle est illustrée dans la figure. 5.

a- Restriction KpnI

Tout d'abord, une grande quantité du vecteur accepteur a été obtenue à partir d'une culture de 50 mL de XL1Blue MRF' transformée avec le vecteur accepteur de la banque. Le plasmide a été

extrait par le kit commercial NucleoSpin® Plasmid de Macherey Nagel. Ceci nous a permis d'avoir une solution de 800 µL contenant 82 µg de plasmide au final.

Le plasmide a été en premier lieu digéré par l'enzyme de restriction KpnI (Fig. 5.A) : 800 µL de la solution de plasmide ont été mélangés avec 2.500 U de l'enzyme KpnI (50 µL d'une solution de KpnI à 50.000 U/mL), 100 µL du tampon 10 mM Bis-Tris-Propane-HCl, 10 mM MgCl2, 1 mM Dithiothreitol à pH 7 (Tp NEB1), 10 µL de BSA et 40 µL d'H_2O. Après incubation 2h3min à 37°C, le plasmide digéré a été purifié par le kit ExtractII de Macherey Nagel et élué dans 400µL volume total.

b – Biotinylation du vecteur par les casettes Kpn

La particularité de cette procédure est la biotinylation du vecteur qui a été réalisée par l'introduction d'une cassette biotinylée. En effet, deux oligonucléotides ont été synthétisés : un oligonucléotide biotinylé et son complémentaire. Ces derniers vont être incorporés au niveau du site KpnI du plasmide digéré.

Les oligonucléotides utilisés sont (Fig. 5.B) :

- ***BiotLinkKpn*** *:* ACAGACAGGGTAC *modifié du coté 5' par greffage d'un groupement biotine.*
- ***LinkKpnRev*** *:* CCTGTCTGT *non modifié.*

100 µL de chaque oligonucléotide ont été mélangés en présence de 20 µL de tampon de ligase, incubés 10 min à 55°C puis refroidis à 4°C pour permettre l'hybridation et la formation de la cassette biotinylée (Fig. 5.C). 220 µL de cette cassette ont été mélangés par la suite avec 300 µL du vecteur digéré avec KpnI, 60 µL de tampon de ligase et 8000 U de ligase (20 µL d'une solution à 400.000 cohesive end units/mL). La ligation a été réalisée o.n. à 16°C.

La quantité de cassette utilisée a été calculée pour avoir un fort excès (400 fois) de celle-ci par rapport au vecteur évitant refermeture du plasmide et favorisant l'incorporation des cassettes biotinylées. Le vecteur biotinylé ainsi obtenu est purifié par le kit ExtractII (Fig. 5.D).

c – Obtention du vecteur semi-biotinylé

L'insertion des motifs nécessite que l'une des extrémités du vecteur soit libérée. Ainsi, une digestion BbsI permettra d'obtenir du vecteur semi-biotinylé prêt à intégrer les motifs.

720 µL du vecteur biotinylé ont été mélangés avec 100 U de BbsI (20 µL d'une solution d'enzyme à 5.000U/mL), 90 µL de tampon 10 mM Tris-HCl, 50 mM NaCl, 10 mM MgCl2, 1 mM Dithiothreitol pH 7.9 (Tampon NEB2) et 70 µL d'H_2O. La digestion a été réalisée 2 h à 37°C

puis le produit de digestion a été purifié : ***Plasmide semi-biotinylé*** *prêt pour l'insertion des motifs (Fig. 5.E).*

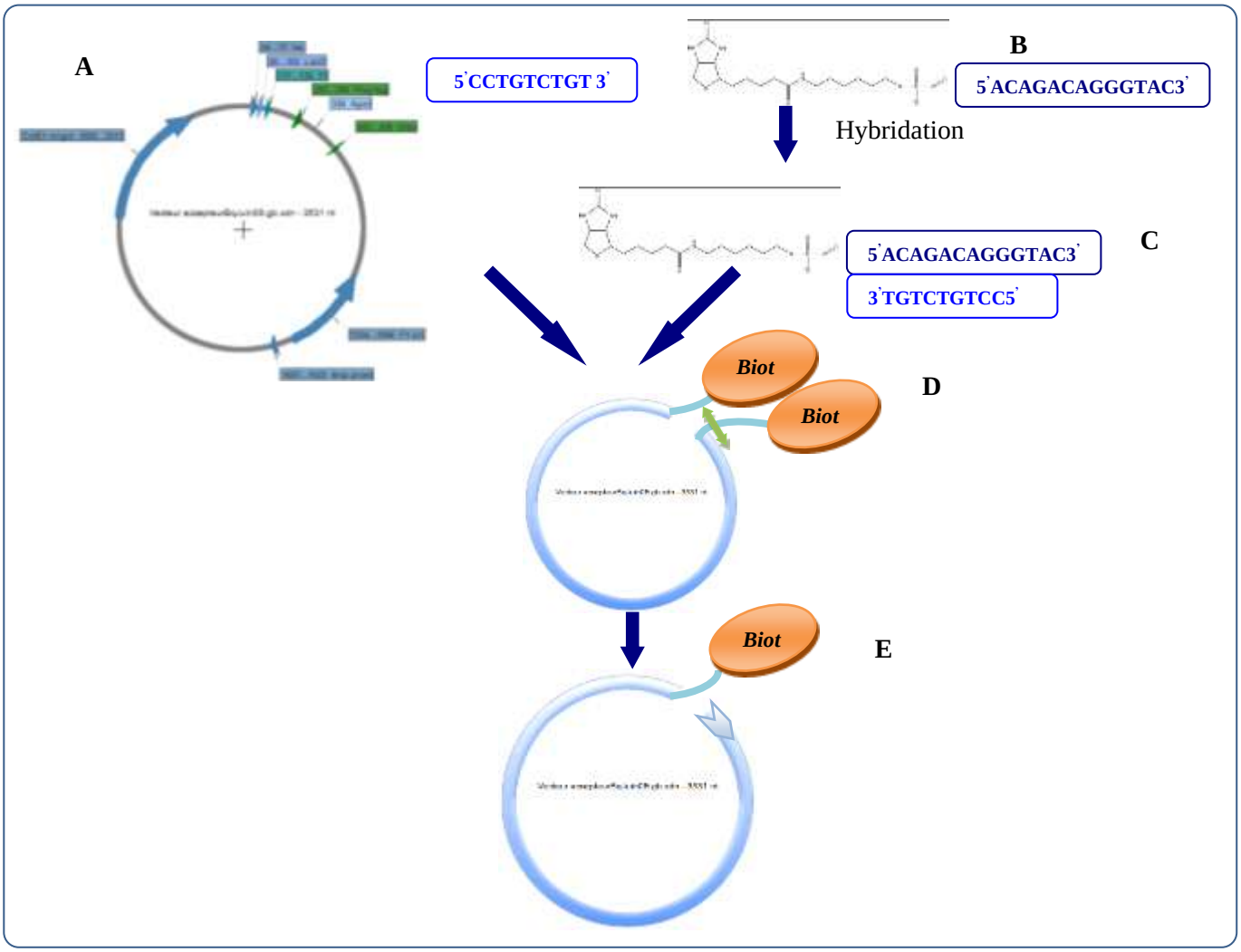


Fig. 5 : Schéma général de la procédure suivie pour l'obtention du vecteur accepteur semi-biotinylé : A : Le vecteur accepteur de la banque (Section A) digéré avec KpnI. B : Les oligonucléotides synthétisés dont l'hybridation a permis d'obtenir la casette Kpn biotinylées (C). D : Plasmide biotinylé obtenue par la ligation du vecteur (A) avec les casettes (C). Ce plasmide est digéré avec BbsI ce qui nous a permis d'avoir le plasmide semi-biotinylé (E).

Une électrophorèse sur gel d'agarose 1% de 2 µL des différentes étapes nous a permis de vérifier la qualité du vecteur à l'issus de chaque étape (Fig. 6).

Le contrôle sur gel réalisé montre qu'une partie de la population du vecteur ne reste pas sous une forme monomérique linéaire (puits C) lors de la ligation des cassettes biotinylées au site Kpn (passage du puits B au puits C). Cela résulte probablement que la quantité de cassette utilisée bien qu'en excès n'est pas encore suffisamment concentrée pour éliminer totalement les autres réactions de ligation en compétition cinétique avec la fixation du *linker* biotinylé. Ainsi, une partie des plasmides parviennent à circulariser et/ou « oligomériser ».

Après vérification de la qualité du vecteur semi-biotinylé, nous avons quantifié la solution que nous avons obtenue : 18 µg de plasmide prêt pour l'insertion des motifs

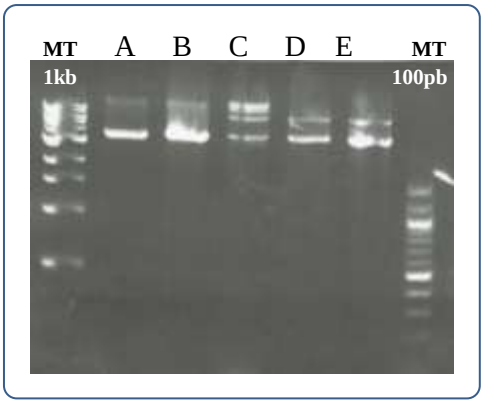


Fig. 6 : Electrophorèse des différentes étapes suivies pour l'obtention du vecteur semi-biotinylé : A : Echantillon de la digestion du vecteur accepteur avec l'enzyme de restriction KpnI. B : Le vecteur accepteur digéré KpnI purifié. C : Le produit de la ligation du vecteur avec les casettes Kpn biotinylées. D : Le vecteur biotinylé coupé par l'enzyme BbsI. E : Le vecteur semi-biotinylé purifié.

1.4. Ligation vecteur - motifs

A ce niveau, nous avons un vecteur dont l'une des extrémités (coté N-terminal de la protéine) est bloquée par la biotine et l'autre extrémité (coté C-terminal) est libre avec l'extrémité libérée par le site BbsI, extrémité qui est cohésive avec celles libérées par le site BsmBI des motifs à insérer.

La ligation est réalisée avec un rapport molaire Fragment / Vecteur de 20. En effet, les 18 µg de vecteur semi-biotinylé ont été mélangés avec 15 µg de motifs préalablement préparé, 800 U de ligase (2 µL d'une solution à 400.000 cohesive end units/mL) et 62 µL de tampon de ligase dans un volume réactionnel total de 620 µL. La ligation a été réalisée o.n. à 16°C. Le produit obtenu est le vecteur avec une extrémité comprenant le motif Ccap, un nombre variable de motifs et une extrémité contenant la moitié constante du motif Ncap et bloquée à la biotine (Fig. 7.A)

1.5. Ajout du motif *Ncap*

A ce niveau, la partie manquante à la séquence avant la fermeture du plasmide, est la moitié variable du *Ncap*. En effet, le *Ncap* de la banque consiste en un motif cap d'une protéine naturelle. Il ressemble aux motifs internes et adopte le même repliement : il comporte une région constante et des positions qui sont rendues aléatoires structuralement équivalentes aux positions 18, 19, 22, 23, 26 et 30 des motifs internes. Une différence majeure est la séquence de la région constante du *Ncap* qui est adaptée à l'exposition au solvant et non pas impliquée dans les interactions intramoléculaire comme chez les motifs internes. Cette partie du *Ncap*, qui est commune à toutes les protéines de la banque, a été intégrée dans la séquence

du vecteur accepteur de la banque. Alors que la région rendue variable et participant à la formation de la surface potentielle d'interaction, est formée par hybridation d'oligonucléotides synthétiques. Ces gènes vont coder pour les positions hypervariables de façon à mimer la diversité naturelle adaptée pour les mêmes positions des motifs internes.

Pratiquement, l'intégration de cette moitié du *Ncap* dans le motif se fait en deux étapes :

- formation des cassettes *Ncap* par hybridation d'oligonucléotides synthétiques.
- Ligation avec les vecteurs obtenus au préalable.

a - Préparation des cassettes stopN-lib / stopN-lib-rev

A ce niveau, le vecteur accepteur contient le *Ccap* sur lequel ont été polymérisé un nombre variable de motif et à l'autre extrémité du vecteur se trouve la moitié constante du *Ncap*. Il manque alors la moitié du motif *Ncap* contenant les positions hypervariables pour pouvoir refermer le vecteur et obtenir la bibliothèque.

Des oligonucléotides conçus pour coder pour les positions conservées et les positions hypervariables de cette moitié du Ncap ont été synthétisés chez MWG eurofins. 25 µL de chaque oligonucléotide à 100 µM (for et rev) ont été mélangés en présence de 5 µL de tampon ligase, chauffés à 95°C puis ramenés lentement à 4°C favorisant l'hybridation des paires spécifiques.

b - Capping

Une fois, les moitiés *Ncap* constituées par hybridation des oligonucléotides, elles sont liguées au niveau de l'extrémité compatible du dernier motif intégré.

50 µL des cassettes Ncap obtenues ont été ajoutés au produit de ligation du vecteur avec les motifs avec 800 U de ligase (2 µL d'une solution à 400.000 cohesive end units/mL), 20 µL de tampon ligase et 130 µL d'H_2O. Le mélange est incubé o.n. à 16°C (Fig. 7.B).

1.6. Purification et refermeture des plasmides

La présence de la biotine à l'une des extrémités du vecteur va permettre la capture des plasmides sur des billes magnétiques recouvertes de streptavidine.

Les billes utilisées sont des Dynabeads® MyOne™ Streptavidin C1 (2 mg) , elles ont été tout d'abord lavées dans 1 mL H_2O puis 2 fois dans 1mL de tampon de lavage et d'interaction ; 5 mM Tris, 1 M NaCl et 0.5 mM EDTA (B&W buffer). Par la suite, elles ont été reprises dans 820 µL de B&W buffer (2x) auxquels on a ajouté les 820 µL de la ligation précédente. La fixation du vecteur

semi-biotinylé a été réalisée 30 min à température ambiante avec agitation. L'excès de motif et le milieu réactionnel ont été éliminés par 2 lavages avec un 1mL de B&W buffer et un dernier lavage au tampon 50 mM Tris-HCl, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM Dithiothreitol à pH 7.9 (Tampon NEB3). Par la suite, les billes ont été reprises dans 300 µL de ce tampon, et l'élution a été réalisée par digestion BspMI. En effet, 6 U de BspMI (3 µL d'une solution d'enzyme à 2.000 U/mL) ont été ajoutés à la solution de billes auxquelles les vecteurs biotinylés sont fixés. Après incubation 2 h à 37°C avec agitation, les billes ont été capturées par un aimant et le surnagent contenant le vecteur sans biotine a été récupéré. Les plasmides ont été par la suite purifiés avec le kit Extract II pour être refermés par ligation (Fig. 7.C).

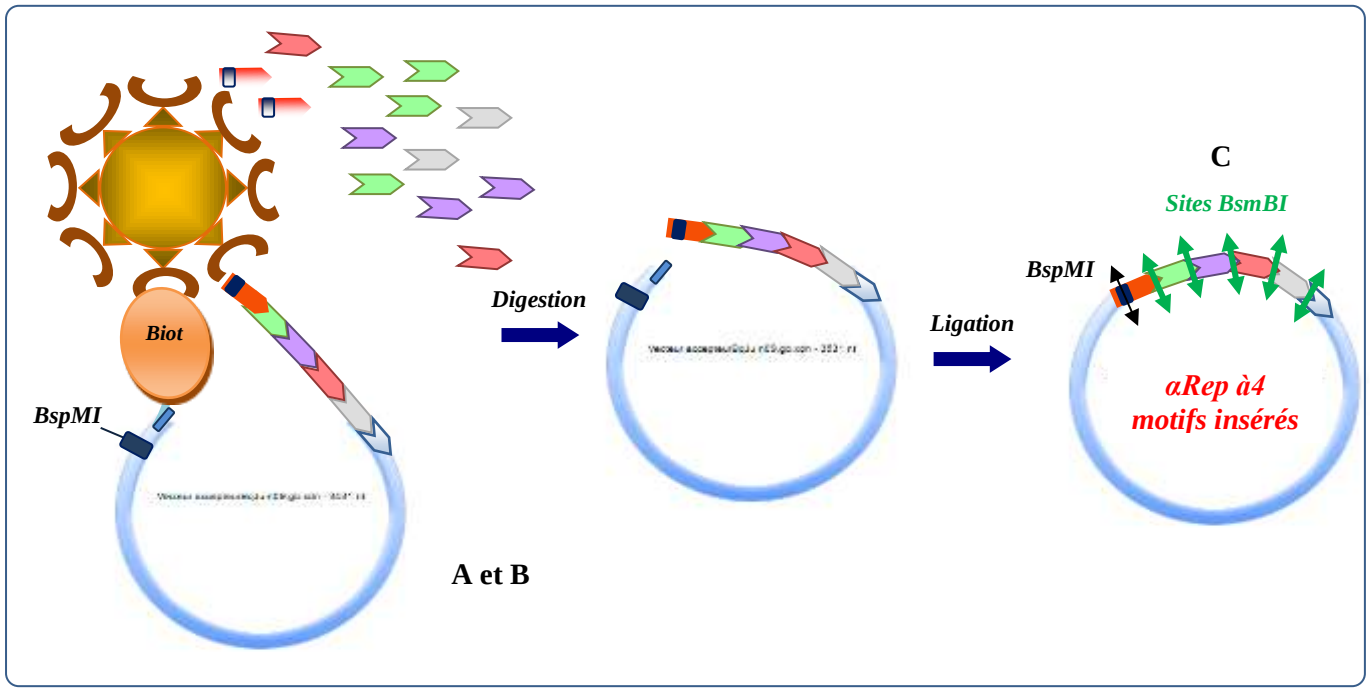


Fig. 7: Schéma illustrant les différentes étapes suivies pour l'obtention des plasmides de la banque. (A) : l'insertion des motifs dans le vecteur accepteur semi biotinylé. (B) : *Capping* : ajout des cassettes de la moitié variable du *Ncap*. (C) : Purification sur billes magnétique couverte de *Strep-Tactin*, élution par la digestion BspMI et fermeture des plasmides par ligation.

1.7. Obtention et caractérisation de la banque d'*αRep* de première génération

a – Obtention de la banque

Les plasmides issus de la purification ont été utilisés pour transformer des XL1Blue MRF' électrocompétentes : une solution finale de plasmide de 75 µL a été obtenue. Chaque échantillon de 100 µl de cellules électrocompétentes commerciales ont été électroporés par 5 µL de plasmides ainsi 15 électroporations ont été réalisées. Suite au choc électrique, les cellules sont immédiatement resuspendues dans 900 µL de milieu Soc et incubées 1 h à 37°C. Toutes les suspensions de bactéries ont été rassemblées et étalées sur 14 boîtes de milieu solide 2YTagar+Amp+Glu. 10 µL de la

suspension ont été utilisés pour le comptage du nombre de variants constituant la banque. Enfin, les bactéries, constituant la banque, ont été raclées des boîtes de milieu solides puis aliquotées et conservées dans du 2YT + Glu 1% + Gly 20%.

Ainsi, la ***banque primaire d'αRep*** a été obtenue. 1 mL de ces bactéries a été utilisé pour extraire un échantillon de plasmide représentatif de cette banque. Le comptage a révélé que celle-ci est constituée de $2.6*10^7$ clones indépendants.

b – Caractérisation de la banque primaire

Une caractérisation de la banque obtenue a été réalisée. Il s'agissait d'évaluer la distribution de la longueur des protéines au sein de cette banque ainsi que d'évaluer la qualité des séquences de clones choisis au hasard.

La distribution du nombre de modules insérés peut être évaluée par une restriction effectuée sur une préparation de plasmide purifiée à partir d'un échantillon collectif de bactéries constituant la banque. Tous les plasmides doivent avoir la même structure générale mais diffèrent par le nombre de modules insérés. Une restriction par les enzymes dont les sites sont situés aux deux extrémités de la séquence des *αRep* insérées (NheI-HindIII) va permettre alors de visualiser la distribution du nombre de motifs insérés (Fig. 8).

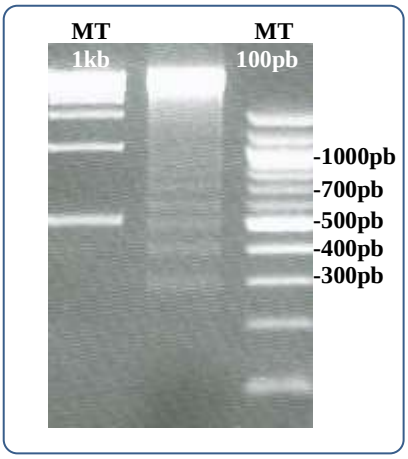


Fig. 8 : Electrophorèse du produit de digestion NdeI-HindIII des plasmides de la banque primaire : Ce gel montre une distribution très régulière du nombre de motifs insérés dans les protéines : de 0 motifs jusqu'à des protéines à 7et 8 motifs.

Le profil de digestion montre clairement une distribution distincte de séquences de différentes tailles. Ces différentes tailles correspondent aux nombres variables de motifs qui ont été intégrés. Nous observons clairement que cette banque comporte des protéines ayant intégrés entre 0 et 8 motifs avec une majorité de protéine comportant entre 1et 5 motifs.

Dans le but d'étudier la proportion des protéines codantes de cette banque ainsi que la fréquence de chaque acide aminé aux positions hypervariables par comparaison aux fréquences codées, nous avons extrait le plasmide de 42 clones, pris au hasard. Ces plasmides ont été par la suite digérés NdeI-HindIII pour déterminer le nombre de motifs insérés (Fig. 9).

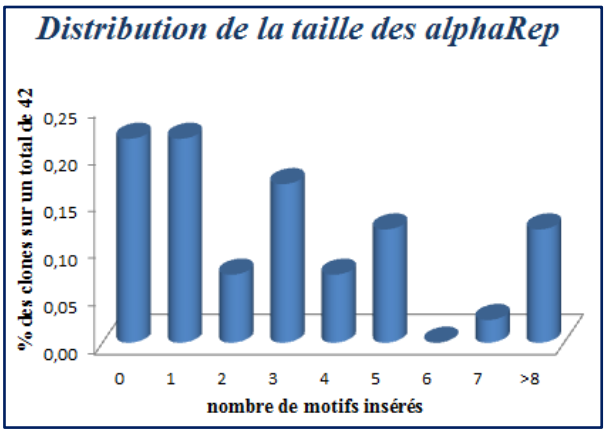


Fig. 9 : Histogramme représentant le pourcentage des clones en fonction du nombre de motifs insérés.

Cet histogramme résume la distribution des tailles des *αRep* dans la banque primaire et confirme ce qui a été observé sur le profil de digestion NdeI-HindIII des plasmides représentant de la banque. En effet, une grande majorité des clones avaient entre 0 et 5 motifs insérés mais il y a aussi 12% des clones qui contiennent au-delà de 8 motifs insérés. 24 clones ont été séquencés : l'analyse des séquences obtenues nous a révélé que 50% des variants ont une séquence avec un codon non-sens ce qui est plus important que les 30% obtenus pour la banque de première génération. L'analyse de la séquence des 73 motifs extraits de ces clones a révélé que 82% de ces motifs insérés sont codants un pourcentage meilleur que celui obtenu dans la banque de première génération (60%).

Nous notons ainsi une amélioration dans la proportion des *αRep* ayant une bonne séquence qui est très probablement due à l'utilisation des oligonucléotides doubles brins pour la construction des motifs qui ont été par la suite amplifiés par RCA. Nous avons ainsi atteint un premier but d'augmenter la fraction des clones codants de la banque primaire. Il reste tout de même un pourcentage non négligeable de motifs avec un codon non-sens qui sont probablement dus à la synthèse des oligonucléotides et leur hybridation. Les limitations issues de ces paramètres que nous ne contrôlons pas vont être contournées par la filtration de la banque par *phage display*.

Nous avons également analysé la distribution des acides aminés aux positions hypervariables à partir de la séquence des motifs extraits et nous les avons comparés avec la

distribution des acides aminés à ces positions dans la collection des *repeat* naturels ainsi qu'avec la distribution que nous avions cherché à obtenir lors de la construction des cercles. La distribution séquencée des acides aminés aux positions hypervariables obtenue dans la banque primaire reflète clairement la distribution codée ce qui suggère que la construction de la bibliothèque n'a pas introduit de biais trop marqué (voir Figure. 17 pages 190-191-192).

Par la suite, il nous paraissait intéressant de tester l'expression en milieu liquide de clones repiqués au hasard comme décrit dans la partie (A-II-4).

La souche d'expression utilisée dans ce test est la souche Rosetta Blue d'E.coli. Les protéines αRep exprimées et transférées sur membrane de nitrocellulose ont été révélées à l'aide d'un anticorps anti-His-tag couplé à la peroxydase.

Le test d'expression en milieu liquide nous a révélé que tous les clones codants extraits de cette banque s'expriment sous forme soluble.

Il est important de mentionner que $2.7*10^7$ clones indépendants constituent une banque relativement restreinte par comparaison aux banques qui peuvent être construites. De plus, 50% seulement des variants sont codants. L'étape suivante a été alors d'éliminer les variants qui comprennent des codons non-sens dans leurs séquences. La taille est cependant suffisante pour échantillonner quasi-exhaustivement tous les modules individuels codés par le schéma de diversification introduit.

2. Filtration de la banque primaire : Banque Filtrée

Dans le but de récupérer la sous population de clones codants de la banque primaire, et par la suite les motifs corrects qui les composent, ces variants ont été collectivement sélectionnés par le biais de l'étiquette *Flag* que les protéines possèdent à leur extrémité N-terminale. La filtration a été réalisée après exposition sur phages des variants de la banque. Ces derniers ont été par la suite dialysés et mis en interaction avec un anticorps anti-*Flag-tag* immobilisé dans des immunotubes.

2.1. Exposition de la banque sur phage

La banque primaire a été conservée par aliquots de 1mL à DO = 74.5. 330 µL de cette suspension, représentant 50 fois la diversité de la banque, ont été utilisés pour ensemencer 250 mL de 2YT+Amp+Tet à une DOi = 0.1. Les bactéries ont été par la suite incubées à 37°C et 220 rpm. A DO = 0.8, les bactéries ont été infectées avec le phage helper (Phaberge) : ajout de 5.7 mL d'une solution $3.5*10^{11}$cfu/mL. Le volume de phage helper a été calculé pour parvenir à une multiplicité de

20. La production de phages a été réalisée comme décrit au préalable. 300 mL de surnagent de culture ont été par la suite récupérés et dialysés : 10 mL de surnageant ont été dialysés dans 1 L de TBS dans 2 bains de dialyse successifs (le 1er 4 h et le 2ème 15 h à 4°C).

Une estimation des particules de phages produits peut être réalisée par une mesure de la DO à 269 nm et un calcul selon la formule :

(DO_{269}*6*10^{16}) / (Nombre de pairs de base du phagemide)

Ainsi, nous avons estimé que la solution de phages dialysés utilisée pour la filtration contient $8.83*10^{12}$ particules/mL (une DO_{269} = 0.53).

2.2. Filtration de la banque primaire

La filtration a été faite par le biais de l'anticorps anti-*Flag-tag*, immobilisé dans des immunotubes (concentration de la solution d'absorption 10 µg/mL). Dans le but d'être sûr de récupérer tous les phages codants de la banque et restreindre au maximum la perte de la diversité initiale de la banque, on a utilisé 5 immunotubes pour la filtration.

Ces 5 immunotubes ont été coatés o.n. avec l'anticorps anti-Flag-tag à 4°C. Par la suite, ils ont été bloqués avec 2 mL d'une solution de TBST BSA 3% : 3 h avec agitation à température ambiante. Une fois les immunotubes préparés, la filtration a été réalisée en incubant 1 mL de phages/immunotube pendant 2 h avec agitation et à température ambiante.

Les phages non spécifiques, qui n'exposent pas à leur surface une αRep repliée, sont éliminés par 15 lavages au TBST suivi de 15 lavages au TBS. Les phages qui sont en interaction avec l'anticorps anti-Flag-tag ont été libérés par élution acide. Nous avons récupéré une solution finale de 4.47 mL de phages élués : 10 µL de cette solution ont été utilisés pour le comptage des phages issus de la filtration alors que le reste a été utilisé pour infecter 50 mL d'une culture de XL1Blue MRF' à DO = 0.7. Les cellules infectées ont été récupérées par centrifugation puis elles ont été reprises dans 5 mL de 2YT et étalées sur 5 boîtes de milieu solide 2YTagar+Amp+Glu.

Les phages de toutes les étapes de la filtration ont été comptés par infection de *XL1Blue MRF'*. Différentes dilutions ont été préparées comme suit :

- pour les phages élués (*Output*) : dilutions 10^{-5}, 10^{-6}, 10^{-7}.
- pour les phages déposés *(Input)*: ***surnageant de culture*** : dilutions 10^{-7}, 10^{-8} et 10^{-9}.
  phages dialysées : dilutions 10^{-7}, 10^{-8} et 10^{-9}.

phages	Dilution	Nombre de colonies	cfu/mL
Input du surnageant de culture	10^{-7}	212	***4.2*10^11***
	10^{-8}	23	
	10^{-9}	2	
Input phages dialysés	10^{-7}	172	***3.72*10^11***
	10^{-8}	20	
	10^{-9}	1	
Elution	10^{-5}	182	***2.2*10^9***
	10^{-6}	5	
	10^{-7}	1	

Les bactéries représentatives des phages élués récupérées sur les grandes boîtes, ont été raclées puis conservées dans du milieu 2YT+Amp+Glu+Gly 20% par des aliquots de 1mL à DO = 136.

Nous avons obtenus une ***banque d'αRep primaire filtrée*** contenant $2*10^9$ clones. Ce nombre, supérieur à la diversité de la banque primaire avant filtration, indique que toutes les séquences codantes de la banque primaire ont une probabilité élevée d'être présente dans la banque filtrée.

2.3. Caractérisation de la banque primaire filtrée

A l'issus de la filtration, nous nous sommes intéressés à la caractérisation de la banque filtrée et en particulier nous avons vérifié si les séquences des clones de cette banque ne contiennent pas de mutations non-sens.

Nous avons extrait les plasmides représentatifs de cette banque filtrée. Par une digestion NdeI-HindII, nous avons comparé la distribution du nombre de motifs insérés dans la banque primaire et dans celle filtrée (Fig. 10).

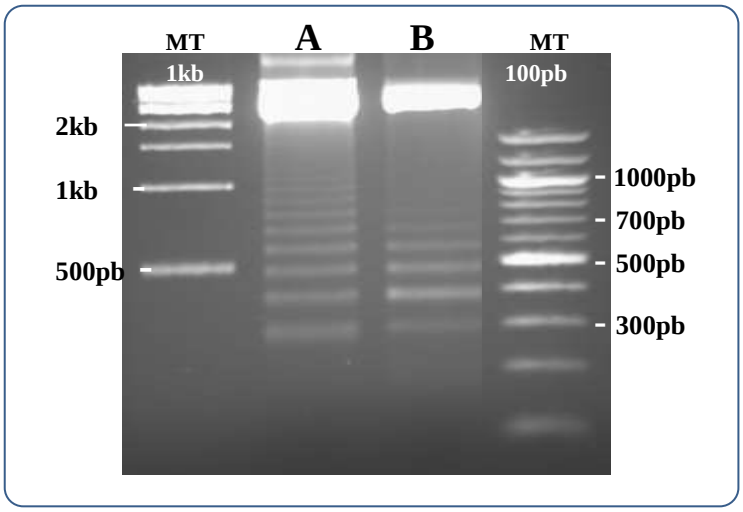


Fig. 10 : Electrophorèse de la digestion NdeI-HindIII des phagemides de la banque primaire (A) et la banque filtrée (B).

Cette digestion montre, comme on pouvait s'y attendre, une diminution dans la fraction des séquences longues. En effet, une protéine à nombre important de motifs a plus de chance d'insérer un motif comprenant un codon non-sens dans sa séquence.

Nous nous sommes intéressés par la suite à l'extraction de plasmides de clones pris au hasard au sein de la banque filtrée, dans le but de les séquencer et étudier la distribution du nombre des motifs insérés et celle des acides aminés aux positions hypervariables. 24 clones ont été repiqués, les plasmides extraits, digérés NdeI-HindIII (Fig. 11) puis séquencés.

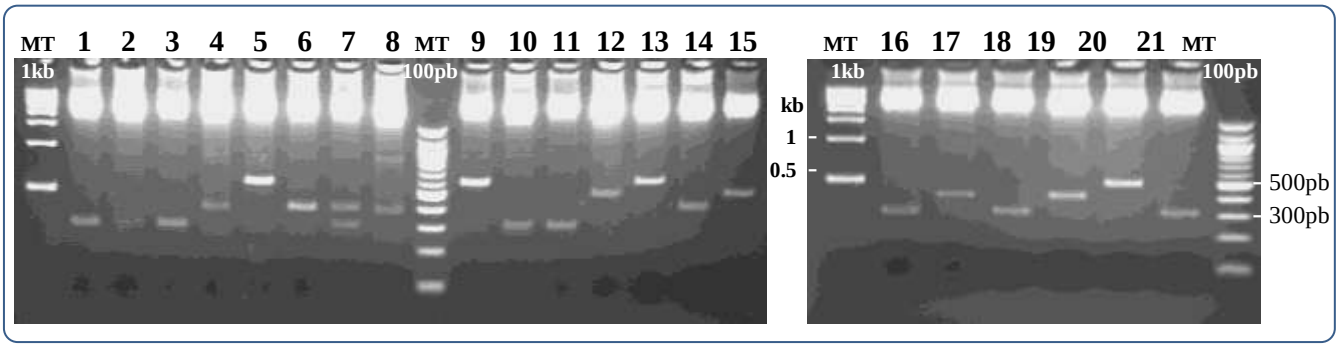


Fig. 11: Profil de digestion NdeI-HindIII des différents plasmides de clones repiqués au hasard de la banque filtrée.

Ce gel confirme une distribution de taille des séquences dans la banque filtrée qui est plus restreinte que celle de la banque primaire. Le diagramme suivant représente cette distribution :

La majorité des clones contiennent entre 0 et 1 motif inséré. Toutefois, il existe certains variants qui ont intégré 2 ou 3 motifs.

Des tests d'expression en milieu liquide ont été réalisés pour ces clones comme décrit au préalable. Ces tests ont montré que les clones ayant intégré au moins un motif s'expriment sous forme soluble.

Il est important de mentionner que tous les clones séquencés (24), de cette banque filtrée, sont codants. Ceci témoigne de l'efficacité de l'approche de filtration par exposition sur phages. Un alignement des motifs insérés a été réalisé (Fig. 12) ce qui nous a permis de

faire une analyse statistique sur les acides aminés par position hypervariable et la comparer à la distribution naturelle celle codée ainsi qu'à celle obtenue lors du séquençage de la banque primaire.

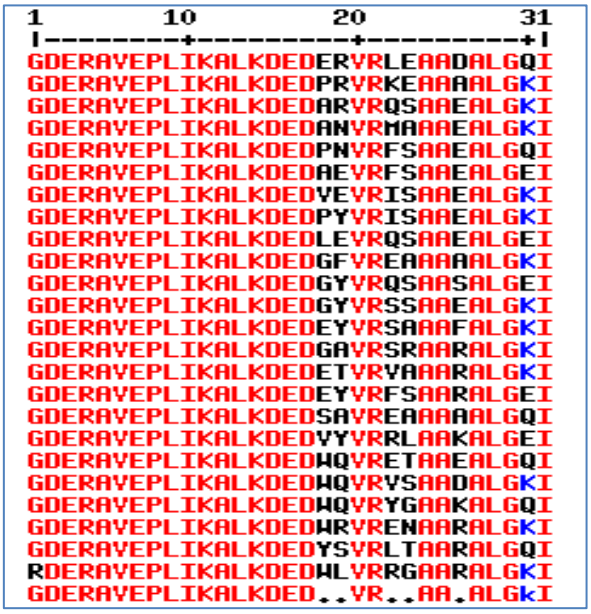


Fig. 12: Alignement des motifs récupérés de la banque filtrée montrant une diversité dans la séquence et l'absence de mutations non-sens.

La filtration a permis d'éliminer les motifs ayant des mutations non-sens et garde une distribution des acides aminés dans les positions hypervariables qui suit la tendance générale codée.

Toutefois, la banque résultante est une banque très restreinte au niveau du nombre de variants et de taille des protéines (une grande majorité entre 0 et 3 motifs). L'étape ultime pour la construction de la banque d'*αRep* de deuxième génération 2.1 serait alors de recréer de la diversité à partir des motifs codants de la banque filtrée et d'augmenter la taille des protéines.

3. « Shuffling » des motifs de la banque filtrée et obtention de la banque de deuxième génération 2.1

La stratégie générale suivie pour obtenir la banque de deuxième génération 2.1 est d'extraire des motifs de la banque filtrée et de les insérer dans un lot de phagemides représentatifs de la même banque dans le but de recréer de la diversité de séquence et obtenir des protéines de plus grande taille.

3.1. Extraction des motifs de la banque filtrée

Dans le but d'extraire les motifs de la banque filtrée, plusieurs approches ont été essayées. Nous avons cherché tout d'abord à extraire les motifs à partir d'une digestion enzymatique des plasmides de la banque puis une purification sur gel d'agarose.

Nous avons tout d'abord extrait une grande quantité des phagemides de la banque filtrée (300 µg d'ADN). Ces derniers ont été digérés avec l'enzyme de restriction BsmBI pour isoler les modules. Le produit de la digestion a été séparé sur gel d'agarose 2%. La bande de 100 pb a été purifiée du gel.

La quantité d'ADN extraite par cette procédure est faible par rapport aux quantités requises. Nous avons alors décidé d'amplifier les motifs codants issus de la banque filtrée par amplification isotherme dans le but d'obtenir les quantités nécessaires pour recréer de la diversité et augmenter la taille des protéines de la banque finale. Nous avions préalablement vérifié que le processus de ré-amplification ne conduisait pas à une fréquence trop importante d'erreurs de séquence (Chapitre II-B-I).

3.2. Amplification des motifs de la banque filtrée

Les motifs extraits de la banque filtrée sont tout d'abord refermés par simple ligation pour obtenir des cercles d'ADN utilisable comme matrice pour l'amplification par RCA.

Les cercles issus de la ligation ont été alors purifiés avec le kit *ExtractII* et par la suite ils ont été utilisés pour réaliser 10 amplifications RCA avec le kit de *TempliPhi*^{TM} *de GE*. Le produit de l'amplification a été digéré partiellement avec l'enzyme de restriction BsmBI (Fig. 13).

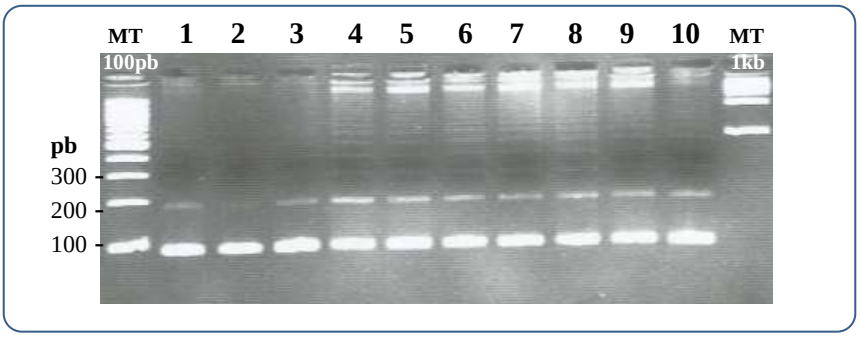


Fig. 13 : Profil de digestion BsmBI des produits des 10 amplifications isothermes réalisées à partir des cercles de motifs extraits de la banque filtrée.

A l'issus de cet amplification, nous avons obtenu une quantité finale de motifs de 33 µg.

3.3. Préparation du vecteur et obtention de la banque finale

L'objectif est de recréer de la diversité dans la banque en mélangeant les motifs et augmentant la taille des séquences. La stratégie générale consiste à digérer les phagemide de la banque filtré avec deux enzymes : une première qui libère les motifs *αRep* (BsmBI) et une deuxième qui coupe le plasmide à 3 sites différents (BsaI).

L'utilisation de la deuxième enzyme vise à retarder les événements de refermeture des plasmides sur eux-mêmes. L'ensemble de la procédure permet de libérer les motifs contenus dans l'échantillon de plasmides, de les mélanger entre eux ainsi qu'avec un ensemble de motifs filtrés et amplifiés. Dans le but de diversifier la taille des séquences, les motifs ont été ajoutés avec différents excès par rapport au plasmide.

60 µg de vecteur de la banque filtrée ont, tout d'abord, été digérés avec les enzymes de restriction BsmBI et BsaI. Ils ont été par la suite purifiés puis ligués en présence des motifs amplifiés. 6 ligations ont été réalisées en variant la quantité de motifs ajoutés tout en gardant le même volume de réaction (Tableau 1).

Rapport Insert/ Vecteur	0	5	10	20	50	100
Vecteur (132.9 ng/µL) (µL)	37.7	37.7	37.7	37.7	37.7	37.7
Inserts (21 ng/µL) (µL)	0	32.8	65.5	132.25	328.1	700 (sol à 20.1 ng/µL)
Ligase (µL)	2	2	2	2	2	2
Tampon ligase (µL)	20	20	20	20	50	82
H_2O	140.2	107.5	74.7	9	82.2	0
Volume total	200	200	200	200	500	820

Les 6 ligations ont été réalisées par incubation o.n. à 16°C. Elles ont été par la suite purifiées par le kit ExtractII sur 6 colonnes et éluées avec 30 µL de tampon d'élution dilué 10 fois. 21 µL de chaque produit de ligation ont été prélevés pour constituer le mélange final de la banque. Nous avons ainsi procédé à 25 électroporations de bactéries XL1BlueMRF' électrocompétentes : 5 µL du mélange de plasmides ligués ont été utilisés pour transformer 100 µL de XL1Blue MRF' électrocompétentes. Les différentes transformations ont été rassemblées (25 mL) : 10 µL ont été utilisés pour le comptage du nombre de variants constituant la banque et le reste a été étalé sur 25 grandes boîtes de milieu solide 2YTagar+Amp+Glu puis incubé o.n. à 37°C.

Nous avons aussi voulu voir l'effet du rapport (vecteur /insert) sur la distribution et la taille des protéines de la banque.

Nous avons alors réalisé 6 autres électroporations avec 5 µL du produit des différentes ligations. Les bactéries transformées ont été étalées sur du milieu solide 2YTagar+Amp+Glu puis incubées o.n. à 37°C. Les tapis de bactéries obtenus sur les 25 grandes boîtes ont été raclés, récupérés dans un volume total de 300 mL de 2YT+Amp+Glu+Gly 20% puis aliquotés par 1.5 mL à DO= 60. De la même façon nous avons récupéré les bactéries issues des transformations des ligations avec les différents ratios (vecteur/ insert).

Nous avons ainsi obtenus la ***banque d'αRep de deuxième génération 2.1***. L'étape suivante était la caractérisation de cette banque ainsi que les différentes ligations.

4. Caractérisation des différentes ligations et de la banque d'αRep de deuxième génération 2.1

4.1. Caractérisation des différentes ligations réalisées

Les bactéries récupérées à l'issus de l'électroporation ont été utilisées pour l'extraction des plasmides correspondant. Ces plasmides ont été par la suite digérés avec les deux enzymes de restriction NdeI-HindIII. Le produit de digestion a été séparé sur gel d'agarose 2 % (Fig. 14).

Nous visualisons sur ce profil que l'ajout des motifs à différents rapport [Vecteur/Insert] en gardant un même volume réactionnel a permis d'augmenter largement la gamme de tailles pour les séquences de la banque. Les séquences les plus longues comportent plus de 12 motifs (séquence d'au-delà de 1.5kb).

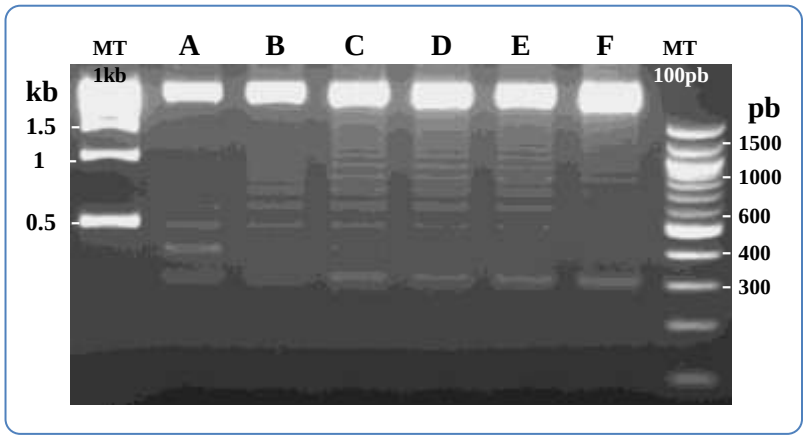


Fig. 14: Profil de digestion NdeI-HindII des vecteurs issus des différents types de ligation :
A : rapport Insert/ Vecteur = 0. B : rapport Insert/ Vecteur = 5. C : rapport Insert/ Vecteur =10.
D : rapport Insert/ Vecteur = 20. E : rapport Insert/ Vecteur = 50. F : rapport Insert/ Vecteur = 100.

Nous remarquons aussi que la bande de 300 pb qui correspond aux variants n'ayant intégré aucun motif (mais seulement le *Ncap* et *Ccap)* n'a pas diminué même en présence d'un grand excès en motifs par rapport au vecteur. Ceci peut être expliqué par l'absence du site BsmBI nécessaire à l'intégration des motifs. En effet lors de l'étape de *shuffling*, les plasmides de la banque sont digérés BsmBI et BsaI avant d'être ligués en présence d'excès de motifs mais les plasmides qui ne contiennent que le *Ncap-Ccap* ne contiennent pas le site BsmB1 et ne sont effectivement digérés que par la BsaI (Fig. 15). Ils sont alors coupés en 3 sites libérant 4 fragments d'ADN qui sont par la suite refermés pour reconstituer un plasmide contenant toujours le *Ncap-Ccap*.

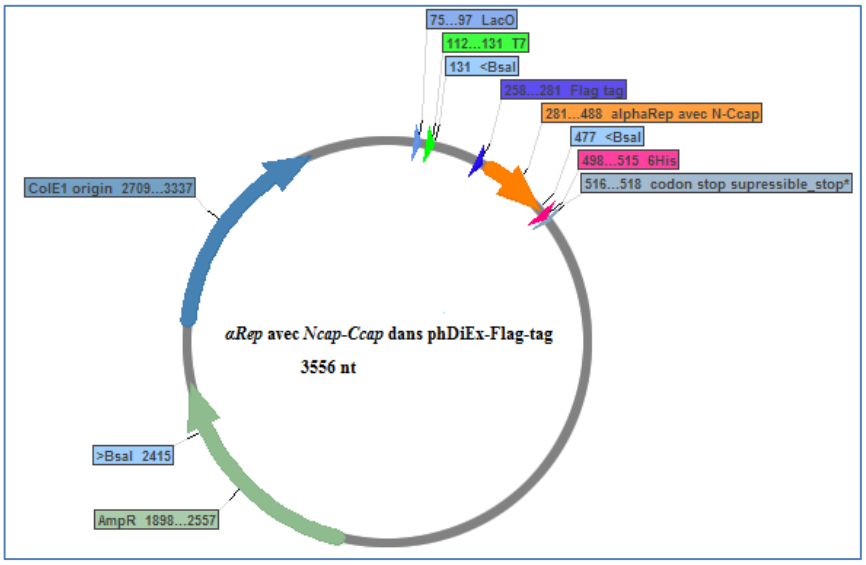


Fig.15 : Carte d'un plasmide contenant une *αRep* n'ayant qu'un *Ncap-Ccap* montrant 3 sites BsaI et aucun site BsmBI qui sert à l'intégration des motifs.

L'étape suivante consiste à vérifier si la banque comporte la même distribution des tailles des protéines.

4.2. Caractérisation de la banque d'*αRep* de deuxième génération 2.1

Le comptage des variants de cette banque nous a permis de déterminer qu'elle est constituée de $1.7*10^9$ clones indépendants. Ce qui est plus important que le nombre de variant de la première banque. Un lot de bactéries représentatives de la banque a été utilisé pour extraire les phagemides. Ces derniers ont été digérés NdeI-Hind III (Fig. 16).

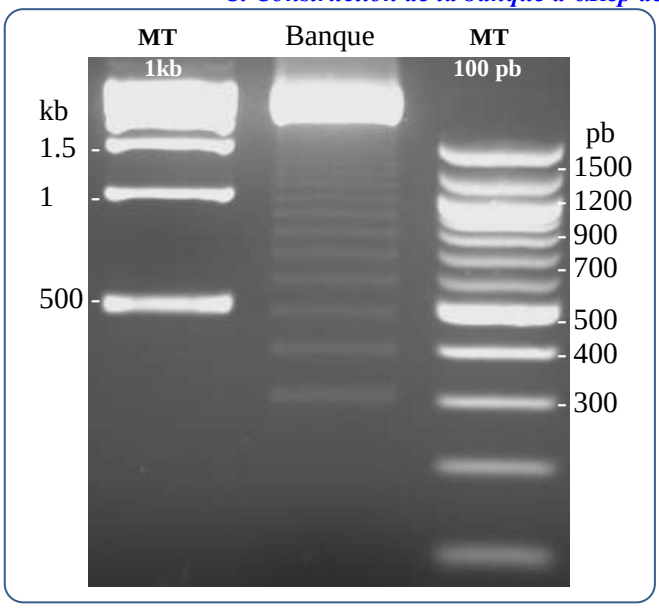
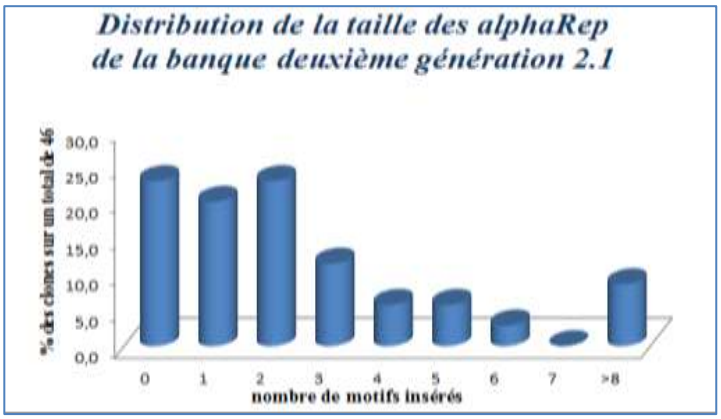


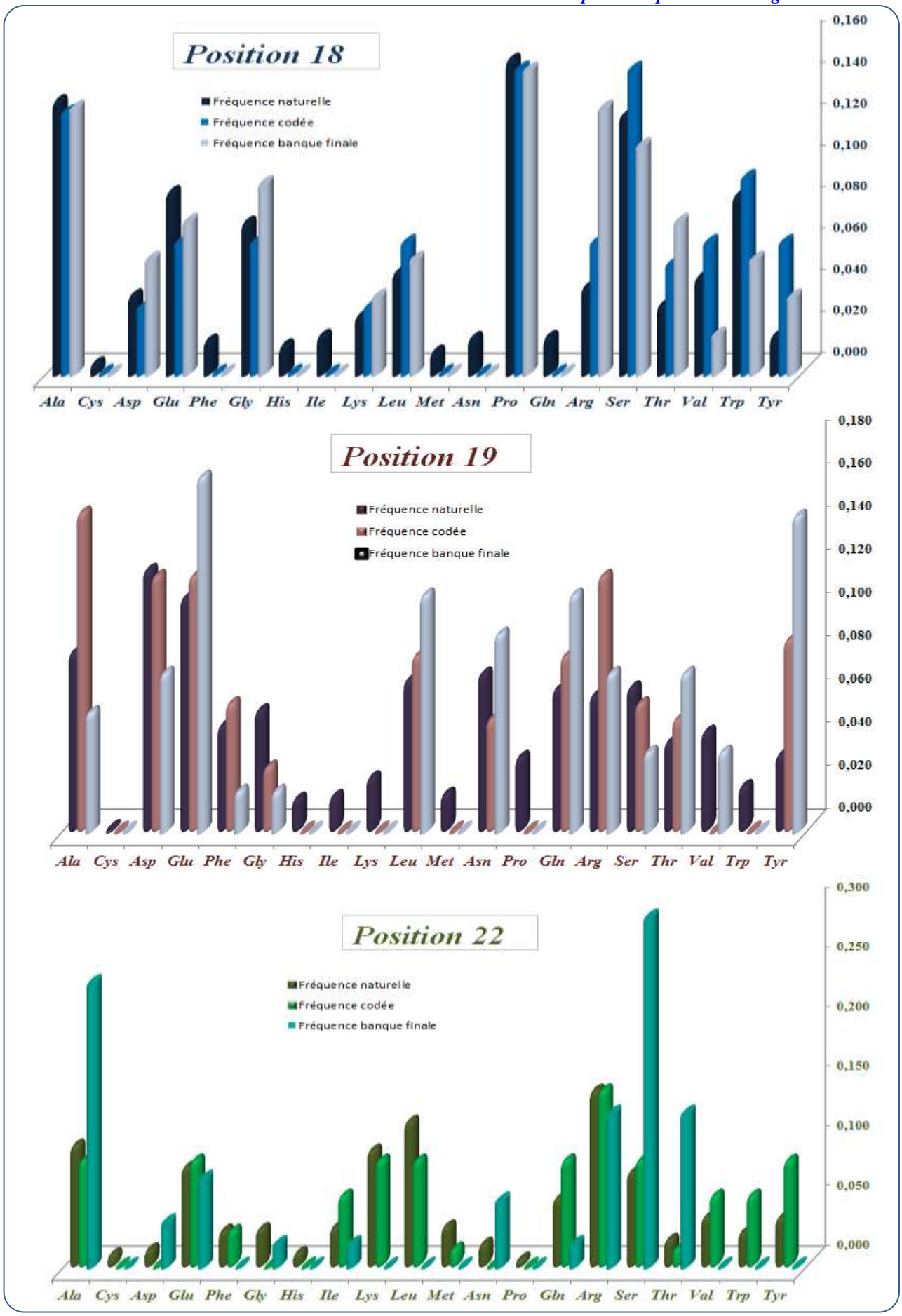
Fig. 16 : Profil de digestion NdeI-HindIII du lot de vecteur représentatif de la banque d'*αRep* de deuxième génération 2.1 montrant que les protéines de cette banque peuvent contenir un nombre de motifs insérés au-delà de 13motifs.

46 colonies individuelles, issues des boîtes de comptage, ont été repiquées au hasard et utilisée pour extraire les plasmides. Ces derniers ont été par la suite digérés NdeI-HindIII et séquencés. La distribution des tailles de ces clones montre une grande diversité dans le nombre de motifs insérés : la majorité des clones contient entre 1 et 3 motifs mais il y a même des clones qui contiennent 12 motifs.



Cet histogramme confirme les observations faites au préalable concernant la proportion des *αRep* à 0 insert. En effet, la fraction de cette sous-population d'*αRep* reste quasi constante entre la banque primaire, celle filtrée et la banque finale. Ceci peut être expliqué par l'absence du site BsmBI nécessaire à l'intégration des motifs lors du *Shuffling*.

Les plasmides de ces clones ont été séquencés. Les motifs obtenus ont été alignés puis la distribution des acides aminés aux positions hypervariables (Fig. 17) a été évaluée.


Position 18
Fréquence naturelle
Fréquence codée
Fréquence banque finale
0,160
0,140
0,120
0,100
0,080
0,060
0,040
0,020
0,000
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr
Position 19
Fréquence naturelle
Fréquence codée
Fréquence banque finale
0,180
0,160
0,140
0,120
0,100
0,080
0,060
0,040
0,020
0,000
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr
Position 22
Fréquence naturelle
Fréquence codée
Fréquence banque finale
0,300
0,250
0,200
0,150
0,100
0,050
0,000
Ala Cys Asp Glu Phe Gly His Ile Lys Leu Met Asn Pro Gln Arg Ser Thr Val Trp Tyr

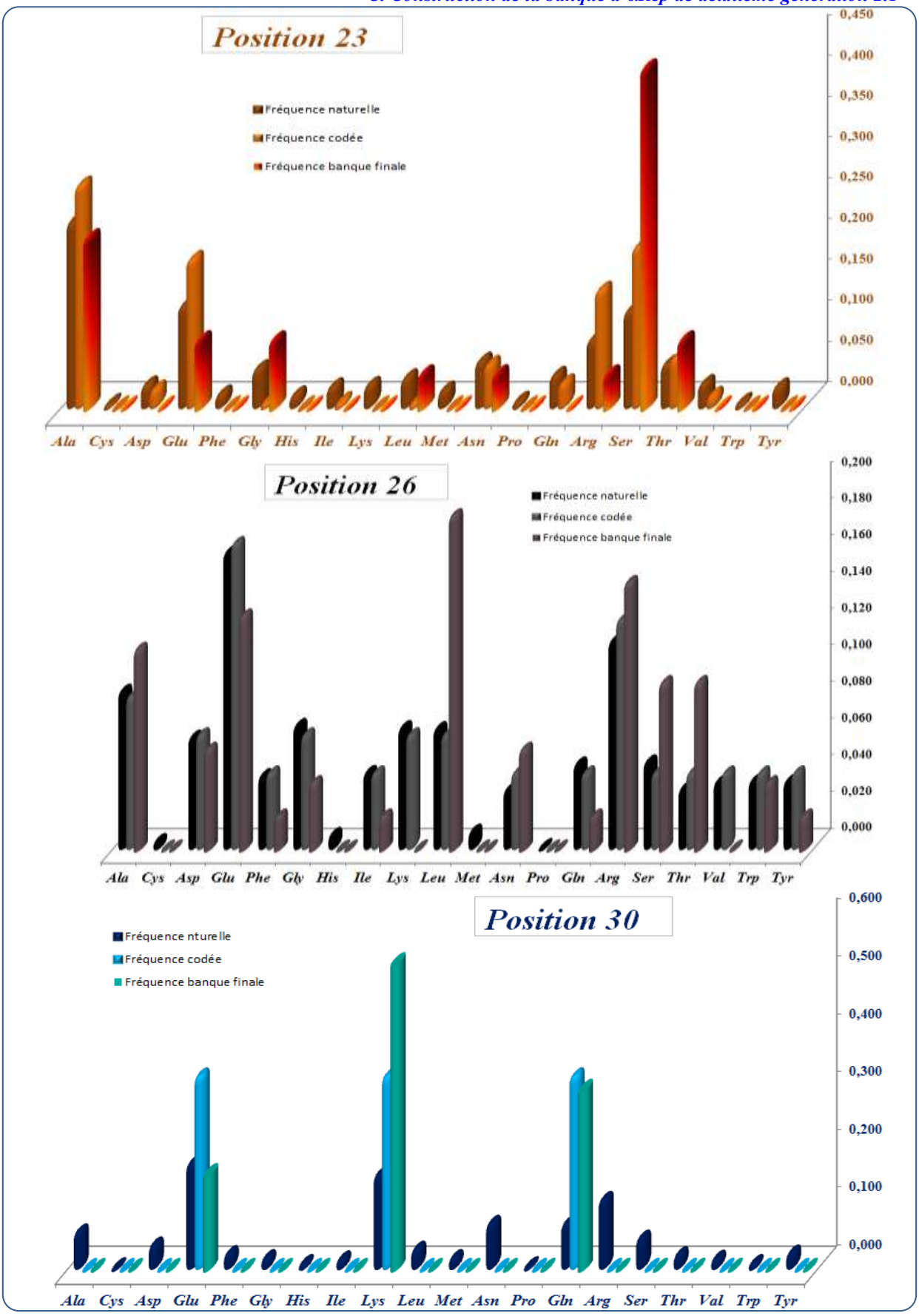


Fig. 17 : Diagramme de la distribution des acides aminés aux positions hypervariables comparant la distribution naturelle avec la distribution codée et celle de la banque de deuxième génération 2.1 (les colonnes correspondantes en allant de la gauche vers la droite).

La distribution des acides aminés aux positions hypervariables dans la banque de deuxième génération 2.1 suit la tendance générale qui a été codée sans apparition de biais particulier pour la position 18. Pour les autres positions bien que la diversité suive le codage escompté il y a certains biais qui apparaissent : Ser et Ala sur-représentés, pas d'aromatiques à la position 22, Ser sur-représenté position 23 et Leu en position 26. Ces biais peuvent être dus à des différences liés à la synthèse des motifs dégénérés par exemple à la surreprésentation d'un oligonucléotide particulier, ou encore ou à un biais d'amplification. Ces biais peuvent aussi résulter en partie d'un l'échantillonnage insuffisant.

L'analyse de séquence a révélé aussi une autre anomalie: certains variants voient une partie de leur *Ncap* manquée à la séquence (Fig. 18).

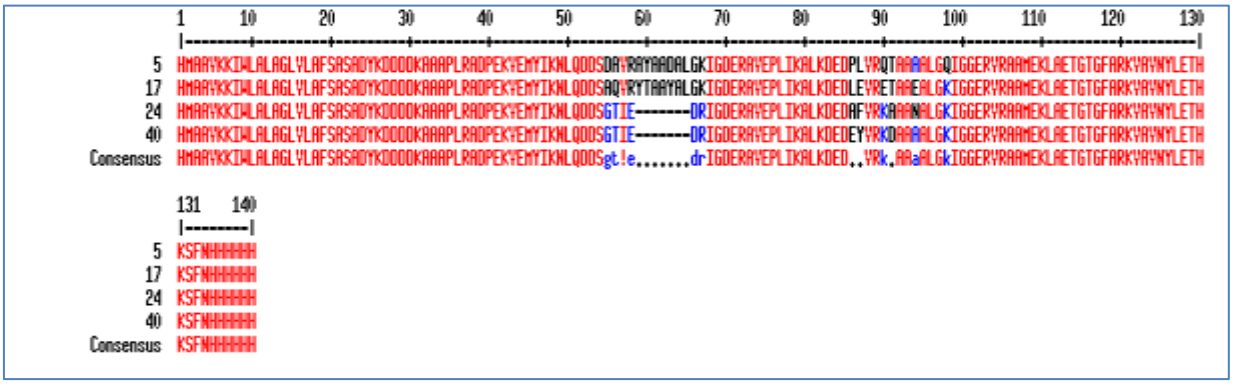


Fig. 18 : Alignement des séquences de 4 *αRep* différentes de la banque de deuxième génération 2.1 montrant la séquence entre AA_{22} et AA_{29} qui manque pour 2 variants parmi 4.

Cette séquence manquante correspond à la moitié rendue variable du *Ncap*. Ceci peut être dû à une quantité de cassettes *Ncap* insuffisante lors de la construction de la banque primaire. En effet, les cassettes étant pas en excès vont donner des plasmides qui vont se refermer sur eux même après la digestion BspMI sans qu'elles y soient intégrées.

Globalement la caractérisation de la banque après filtration et *shuffling* montre que les séquences comportent une distribution de longueur, que la plus grande fraction des séquences est effectivement codante et qu'une diversité réelle des chaines latérales est observée aux positions hypervariables. Cette diversité suit globalement la distribution naturellement observée et en particulier exclue les chaines latérales défavorables à la structure du motif.

Un paramètre aussi important que la séquence des variants *αRep* est l'expression des protéines et leur solubilité que nous avons ensuite évaluées par *Cofiblot* (test de solubilité).

Comme décrit au préalable (B-II), des bactéries Rosetta Blue ont été alors transformées avec les phagemides représentatifs de la banque. Différentes dilutions ont été réalisées pour obtenir des colonies séparées. Deux plaques 96 puits ont été préparées puis utilisées pour ce test. L'expression

des protéines a été révélée par un anticorps primaire anti-His-tag suivi d'un anticorps secondaire antisouris fluorescent (Fig. 19).

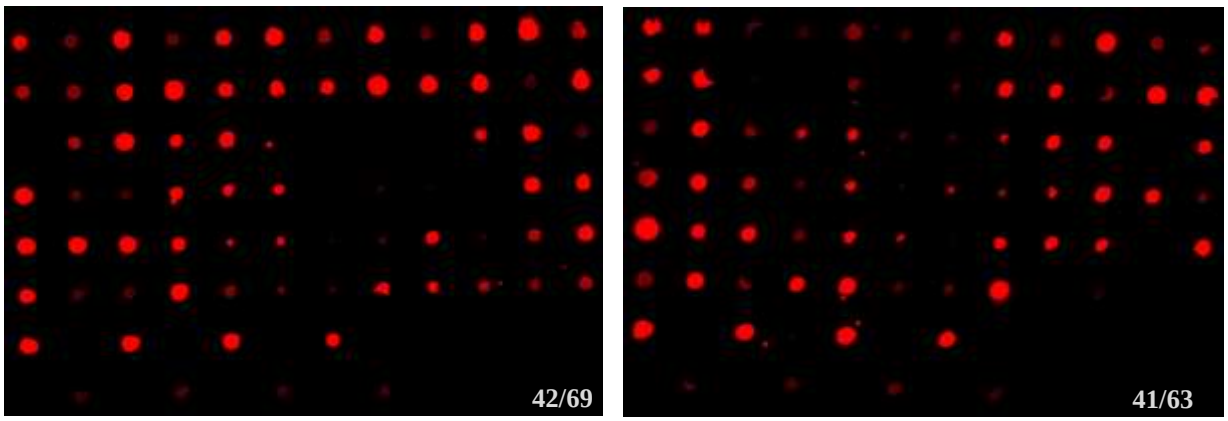


Fig. 19 : Membranes de nitrocellulose révélée de deux *Cofiblot* correspondant à deux plaques de clones de la banque de deuxième génération 2.1. Le témoin positif utilisé est un clone de la banque de deuxième génération 2.0 alors que le témoin négatif est un clone de la même banque dont la séquence comprend un codon non-sens.

Les C*ofiblot* réalisés pour la banque de deuxième génération 2.1 montrent qu'aux alentours de 60 % des clones de cette banque s'expriment sous forme soluble. Les colonies qui ne s'allument pas sont pour partie des colonies qui s'étaient mal développées.

Conclusion

Dans cette partie de la thèse, nous avons exposé la procédure suivie pour la construction de la banque d'*αRep* de deuxième génération 2.1. Cette banque contient aux alentours de **$1.7*10^9$ clones indépendants** qui sont **codants**, **s'expriment sous forme soluble** et dont la séquence, aux positions rendues aléatoirement variables, mime la diversité naturelle. Cette banque se différencie, de celle de première génération, par la **diversité dans la séquence** des positions hypervariables et le **nombre de motifs insérés**. Ces deux niveaux de diversité de cette banque sont les deux critères essentiels pour trouver des variants susceptibles d'établir une large gamme d'interactions avec une variété de molécules cibles.

Peut-on-trouver au sein de cette banque d'αRep de deuxième génération2.1 des variants qui peuvent reconnaître spécifiquement et avec haute affinité des molécules cibles préalablement choisies ?

D. Sélection, identification d'interacteurs pour des molécules cibles et caractérisation des complexes obtenus

Introduction

Les résultats prometteurs obtenus précédemment nous ont encouragés à tester expérimentalement cette banque et à tenter de sélectionner des interacteurs spécifiques de trois protéines cibles choisies *a priori*. Les cibles choisies sont d'une part une cible modèle pour laquelle des interacteurs spécifiques pourraient être pratiquement utiles : la TEV protéase et deux cibles choisies (EbsI et Upf2) dans le but d'identifier des outils de co-cristallisation. Ces deux dernières cibles ont été choisies en collaboration avec l'équipe de H. Van Tilbeurgh (Fonction et Architecture des Assemblages Macromoléculaires : FAAM) au sein de notre institut.

1. Présentation des cibles choisies

1.1. La TEV protéase

La TEV[4] protéase est le nom commun du domaine catalytique de la protéine d'inclusion nucléaire a (Nuclear Inclusion a : NIa) codée par le virus etch du Tabac. C'est un domaine de 27 acides aminés dont l'activité « protéase » est hautement spécifique. Elle reconnaît, en effet, la séquence Glu- X-X-Tyr-X-Gln-(Gly/Ser) et le clivage a lieu entre (Gln/Gly) ou (Gln/Ser). La structure de la TEV protéase, résolue par Phan et ses collaborateurs (Phan, 2002), montre une structure similaire à celle des protéines à Serine. En effet, l'activité est assurée par un « trio catalytique» Ser-Asp-His pour la majorité des protéases à Ser à la différence que la TEV qui contient une Cys nucléophile à la place de la Ser. Grâce à sa spécificité de clivage, la TEV protéase est souvent utilisée comme moyen de clivage des protéines de fusion.

Nous nous sommes intéressés à cette enzyme d'une part parce que c'est une protéine bien caractérisée : les propriétés physico-chimiques de cette protéine sont déterminées. De plus, les interacteurs potentiels à la TEV protéase seront éventuellement utiles comme moyen de séparation de la TEV protéase d'avec les substrats et produits de la réaction de clivage.

1.2. Les protéines choisies comme cible pour générer des outils de cristallogenèse

L'équipe FAAM de notre institut s'intéresse à l'étude structurale et fonctionnelle de complexes macromoléculaires. Grâce à cette collaboration, la structure de l'α-Rep-n4-a été résolue (Urvoas et *al.*, 2010). Nous avons alors constaté que les premiers essais de

[4] http://www.cardiff.ac.uk/biosi/staffinfo/ehrmann/tools/TEVprot.html

cristallogenèse semblaient particulièrement favorables. Ceci nous a conduits à envisager d'utiliser les *αRep* comme des outils « d'aide à la cristallisation ». Dans le cadre de ce projet, je me suis particulièrement intéressée à la sélection d'interacteurs pour deux protéines impliquées dans la voix NMD (*Non sense Medicated Decay*).

La voie NMD est une voix responsable de l'élimination des ARNm portant des codons stop précoces. En effet, la traduction de ces ARNm aberrants va produire des protéines tronquées ayant, potentiellement, des effets dominants néfastes pour l'organisme. Cette voie, intervenant en amont de la synthèse protéique, joue le rôle de détection et d'élimination de ces ARNm avant le début de la traduction. Ce mécanisme s'avère très conservé chez les eucaryotes depuis la levure jusqu'à l'homme. Chez l'homme, 7 protéines ont été décrites essentielles pour le bon fonctionnement de ce mécanisme :

- Upf1, Upf2 et Upf3 : formant le complexe de « surveillance »
- Smg1, Smg5, Smg6 et Smg7 : impliquées dans la phosphorylation/ déphosphorylation d'Upf1 ce qui est essentiel pour le bon déroulement de la voix NMD.

Chez la levure, en revanche, seulement Upf1, Upf2 et Upf3 ont été décrites. C'est dans ce contexte que l'équipe FAAM s'intéresse à l'étude structurale de deux protéines de la levure impliquées dans le mécanisme NMD : EbsI et Upf2.

a- La protéine EbsI

Initialement, EbsI a été décrite comme répresseur de la traduction chez *S.cerevisiae*. Par la suite, l'étude de Luke (Luke et al. 2007) (Luke et *al.*, 2007) a montré que la partie N-terminale d'EbsI contient un domaine, nommé motif 14-3-3 qui est homologue à celui de Smg7. Ils ont démontré aussi que la délétion d'EbsI induit un défaut de la voix NMD. EbsI semble alors être un acteur intéressant à étudier dans cette voix chez la levure.

b - La protéine Upf2

Cette protéine, fait partie du complexe de surveillance au niveau du codon stop précoce. La structure tridimensionnelle des complexes formés par les partenaires Upf1 et Upf3 a été étudiée. Upf2 est une grande protéine qui interagit d'une part avec Upf1 via son domaine C-terminal et d'autre part avec Upf3 via son domaine médian. Toutefois, le domaine N-terminal d'Upf2 reste peu caractérisé structuralement. L'étude que nous menons vise alors à rendre possible l'étude structurale de ce domaine : Upf2 1-360.

2. Sélection par phage display d'interacteurs pour les trois protéines cibles

Le processus de sélection appliqué pour les 3 protéines cibles est le même avec certaines variantes visant l'adaptation du processus aux conditions optimales d'immobilisation des cibles. Un tour de sélection par *phage display* comporte 3 phases essentielles :

- Immobilisation de la cible.
- Mise en contact des phages avec la cible et lavages.
- Elution et récupération des phages spécifiques.

2.1. Production des phages de la banque

C'est une étape très importante. Les phages ont été produits de sorte à couvrir la totalité de la diversité de la banque qui est estimée à $1.7*10^9$ variants.

La banque a été conservée par aliquots de 1 mL à DO = 60. Une culture de 300 mL 2YT+Amp+Tet a été inoculée avec 300 µL de ces bactéries ce qui correspond à une DOi = 0.1 et représente 10 fois la diversité de la banque. La culture a été incubée à 37°C sous agitation à 220 rpm. A DO = 0.6, les bactéries ont été infectées avec 9 mL de phage helper (20 fois la diversité de la banque) : incubation 30 min sans agitation à 37°C puis 30min à 37°C et 80 rpm. Les bactéries infectées ont été collectées par centrifugation (10min, 4000g) puis elles ont été resuspendues dans 300 mL de 2YT+Amp+Kan et incubées o.n. à 30°C pour la production de phages.

Les phages produits et libérés dans le surnageant de culture ont été séparés des bactéries par centrifugation (30 min, 8000g) puis dialysés en deux bains : un premier bain 4 h à 4°C et un deuxième o.n. à 4°C. Les phages destinés pour les sélections contre EbsI et la TEV protéase ont été dialysés dans du TBST alors que ceux utilisés pour les sélections de Upf2 ont été dialysés dans du tampon acétate.

L'utilisation de ce tampon pour la dialyse des phages vise à garder la protéine dans la meilleure conformation car c'est une cible très fragile, difficile à produire et encore plus à cristalliser d'où l'intérêt de lui trouver un interacteur.

2.2. Immobilisation des cibles

Pour ces sélections, nous avons opté pour le système d'immobilisation le plus simple. En effet, les cibles ont été immobilisées par simple d'adsorption à la surface des

immunotubes. Chaque protéine cible a été diluée dans le tampon lui permettant d'être dans la conformation naturelle.

Ainsi, des solutions de protéines cibles à 20 µg/mL ont été préparées. La TEV protéase et EbsI ont été diluées dans un tampon classique d'immobilisation (tampon phosphate 50 Mm pH 7) alors que Upf2 a été diluée dans du tampon acétate (le tampon optimisé pour sa purification).

L'utilisation du tampon acétate au cours de l'immobilisation de la cible Upf2 est une nouvelle procédure visant l'immobilisation de la protéine dans sa conformation naturelle. Il nous est alors nécessaire de vérifier l'immobilisation de cette cible, en particulier que son tampon est non conventionnel. La cible porte une étiquette His_6, elle peut être alors immobilisée dans un immunotube puis révélée par le biais d'un anticorps anti-*His-tag*.

Une solution de la protéine Upf2 a été préparée à une concentration de 20 µg/mL. Un mL de cette solution a été par la suite déposé dans un immunotube et incubé o.n. sous agitation à 4°C. Le tube a été par la suite bloqué avec une solution de TBST BSA 3% pendant 4 h à 4°C. Après 3 lavages au TBST BSA 1%. Nous avons révélé la présence de l'Upf2 par le biais d'un anticorps anti-His-tag couplé à la peroxydase. En effet, le tube bloqué a été incubé avec 1 mL d'une solution TBST avec l'anticorps à la dilution 1/20000. 3 lavages au TBST BSA1 % ont été suivis par la révélation par ajout de 500 µL de substrat soluble. La coloration bleue qui apparaît est neutralisée par ajout de 200 µL d'HCl 1 N. Cette coloration témoigne de la présence d'Upf2 immobilisée à la surface de l'immunotube (Fig. 1).

Fig. 1 : Immunotube vérifiant la procédure suivie pour l'immobilisation d'Upf2.

Pour résumer le protocole final suivi pour l'immobilisation des cibles est :

- *Incubation avec 1 mL d'une solution 20 µg/mL o.n. à 4°C.*
- *3 lavages TBST.*
- *Blocage avec 1 mL TBST BSA 3% 4 h à 4°C.*
- *3 lavages TBST BSA 1%.*

2.3. Mise en contact des phages avec la cible et lavages

Une fois les cibles immobilisées sur les immunotubes. Les phages dialysés ont été mis en contact de la cible 2h à 4°C avec agitation. Par la suite, les phages non spécifiques sont éliminés par 15 lavages avec 1mL de TBST suivis de 10 lavages avec 1 mL de TBS.

2.4. Elution et récupération des phages spécifiques

C'est l'étape marquant la fin d'un tour de sélection et permettant l'obtention du matériel de départ du tour suivant. En effet, après l'élimination des phages non spécifiques ou peu affins, les phages liés aux cibles sont libérés par élution acide.

800 µL de Glycine à pH 2 ont été ajouté à l'immunotube pendant 10 min à température ambiante sous agitation. Ils ont été par la suite récupérés et mélangés à 150 µL de tampon Tris 1 M à pH8. Les phages élués ont été utilisés par la suite pour le comptage des Output (10 µL) et le reste pour infection des XL1Blue MRF'.

a - Comptage des phages

Il est important de déterminer le nombre de phages formant unité (cfu/mL) au début et à la fin de chaque tour. En effet, les phages mis au contact de la cible ont été comptés et ceux qui ont été récupérés dans la solution de Glycine neutralisée au Tris. Ceci va permettre de déterminer de rapport *Output*/*Input* une donnée témoin du bon déroulement des sélections.

Le comptage a été réalisé comme suit :

- préparation d'une culture de XL1Blue MRF' à DO = 0.6 dans du milieu 2YT+Tet

- Préparation des dilutions 10 par 10 des différentes solutions de phages (Input et Output) de 10^{-2} jusqu'à 10^{-9} : 50 µL de phage de la dilution n sont dilué dans 450 µL de 2YT pour obtenir la dilution n+1.

- 50 µL des phages dilués ont été par la suite incubés 15min à 37°C et 80 rmp avec 450 µL de bactéries XL1Blue MRF' à DO = 0.6.

- 50 µL des bactéries infectées par les différentes dilutions de phages ont été étalées sur une boîte de milieu solide 2YT agar+Amp+Glu puis incubées o.n. à 37°C.

Les colonies qui ont poussées sur les boîtes ont été comptées et elles reflètent un nombre de phages infectibles : ***nombre de colonies *10 *20*10^{dil} (cfu/mL).***

b - Récupération des phages du tour : Output du tour n = Input du tour n+1

Le reste des phages, non utilisés pour les comptages, a été utilisé pour infecter 30 mL de bactéries XL1Blue MRF' à DO = 0.6. Les bactéries en contact des phages, ont été incubées 30 min sans agitation à 37°C puis 30 min à 37°C et 80 rpm. Par la suite, elles ont été récupérées par centrifugation (10 min, 4000 g). Le culot de bactéries a été repris dans 1 mL de 2YT puis étalé sur grande boîte de milieu solide 2YTagar+Amp+Tet+Glu et incubé o.n. à 37°C. Les tapis de bactéries ont été raclés, aliquotés puis conservés à DO = 50 à - 80°C.

Les bactéries ainsi conservées ont été utilisée pour produire les phages : Input du tour n+1. Elles sont tout d'abord incubées à DO_i = 0.1 dans un volume de 2YT pour avoir 50 fois plus de bactéries que l'Output en phages calculé au tour n. Après incubation à 37°C et 220 rpm et à DO=0.6, elles ont été infectées par les phages helpers : 30 min à 37°C sans agitation puis 30' à 37°C et 80rpm. Les bactéries infectées ont été récupérées par centrifugation et la production de phages a été réalisée en reprenant le culot bactérien dans 20 mL de 2YT+Amp+Kan o.n. à 30°C et 220 rpm.

Nos sélections ont été réalisées en trois tours. En effet les étapes décrites ci-dessus ont été répétées 3 fois avec quelques modifications dans l'incubation des phages.

Les phages dialysés issus du Tour2 ont été incubés 1 h à température ambiante dans des immunotubes qui ont été juste bloqués à la BSA sans présence de cible. Par la suite, ces phages ont été prélevés et mis directement en contact de la cible. Nous visions la minimisation de la proportion des phages non-spécifiques.

La deuxième modification concerne l'élution du 3^{éme} tour pour EbsI et Upf2: on a réalisé 2 élutions par compétition avec de la protéine cible et 1 élution acide finale.

En effet, après les lavages au TBST et au TBS, on a incubé avec 800 µL de solution 10 µM de la protéine cible pendant 10 min sous agitation. Cette solution est prélevée pour former E1. Une 2^{éme} incubation a été réalisée avec une solution de la même concentration de la cible pendant 1h sous agitation formant ainsi E2. La dernière élution effectuée a été une élution acide E3 permettant de récupérer la totalité des phages.

3. Caractérisation des Tours de sélection et identification des variants interagissant avec les cibles

La caractérisation des tours de sélection se fait avec le comptage des phages récupérés à la sortie du tour n de sélection. Ceci va nous permettre d'être sûr, au début du tour n+1 que le tour n a bien permis d'enrichir en clones reconnaissant la cible. L'enrichissement en variants spécifiques d'un tour à un autre se fait par un test comparatif de l'interaction phages des différents tours-cible ***: test phage Elisa en pool***.

Par la suite, il est important d'identifier des variants isolés ou individuels qui exposés sur phages, reconnaissent la cible par un ***test phage Elisa clonal***.

3.1. Test phage Elisa en pool

Le phage Elisa en pool est un test Elisa basé sur la reconnaissance de l'*αRep* exposée sur phage avec la cible immobilisée sur plaque Elisa. Les phages représentatifs de la banque et des différents tours de la sélection (Tour1, Tour2, et Tour3) sont produits et sont mis en contact avec la cible. La révélation des phages fixés à la cible va permettre de visualiser l'enrichissement en phages spécifique entre les différents tours de sélection.

a - Préparation et plan de la plaque Elisa

La première étape consiste à la préparation de la plaque Elisa par immobilisation de la cible.

En effet, une solution de la cible a été préparée à une concentration de 20µg/mL puis 100µL de cette solution ont été incubés dans la plaque o.n. à 4°C et 350rpm. Les puits des colonnes adjacentes, utilisés comme témoins négatifs, ont été incubés o.n. en présence de PBS. Le plan de la plaque est représenté sur la Fig. 2, ci-dessous.

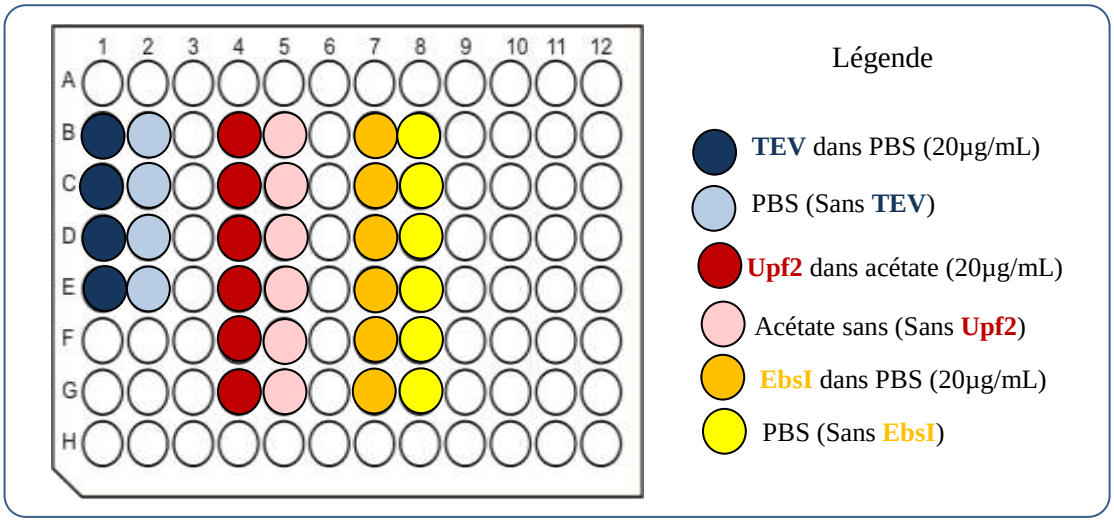


Fig. 2 : Plan de la plaque Elisa préparée pour le test phage Elisa en pool. La ligne B est réservée aux phages issus de la banque. La ligne C est réservée aux phages du Tour1, la D à ceux du Tour2 et les lignes de E vers G pour les phages du Tour 3 issus des différentes élutions (E1, E2 et E3 pour Upf2 et EbsI).

b - Préparation des phages des différents tours de sélection

Les phages représentant la banque, le Tour1 et le Tour2 ont été prélevés à partir des suspensions de phages dialysés utilisées aux différents tours de sélection. Toutefois, les phages

représentant le Output du Tour3 ont été produits à partir des bactéries XL1Blue MRF' récupérées à la fin de la sélection. On obtient alors un échantillon de phages T3 pour la TEV protéase, alors que pour EbsI et Upf2 on a obtenu trois échantillons de phages : T3E1, T3E2 et T3E3. Les phages produits sont par la suite dialysés dans du TBS pour la TEV protéase et EbsI et dans le tampon acétate pour Upf2.

c - Test Elisa et résultats

Après l'immobilisation des cibles, comme présenté au préalable, les puits sont lavés 3 fois au TBST. Par la suite, ils ont été incubés avec 250 µL de TBST BSA 3% pendant 3 h à 15°C. Le blocage a été suivi par 3 lavages au TBST.

Par la suite, 100 µL de chaque échantillon de phage ont été incubés dans les puits avec et sans la protéine cible. L'incubation a été réalisée 2 h à 25°C. Les phages qui n'interagissent pas avec la cible ont été éliminés par 4 lavages au TBST. La présence des phages, reconnaissant la cible immobilisée, a été révélée par incubation 1 h à 20°C avec 100 µL d'une solution de TBST avec l'anticorps anti-M13 couplé à la peroxydase à la dilution 1/5000. Après 3 lavages au TBST, on a ajouté 100 µL de substrat soluble. La réaction du substrat avec la peroxydase donne une coloration bleue qui s'intensifie avec le temps d'incubation. Au bout de 5 min d'incubation sous agitation, la réaction est stoppée par ajout de 100 µL d'acide HCl 1 N. La coloration bleue vire alors à la couleur jaune dont l'intensité est proportionnelle à la quantité d'anticorps anti-M13 ce qui reflète la quantité de phages fixés à la cible. La plaque révélée est représentée dans la Fig. 3.

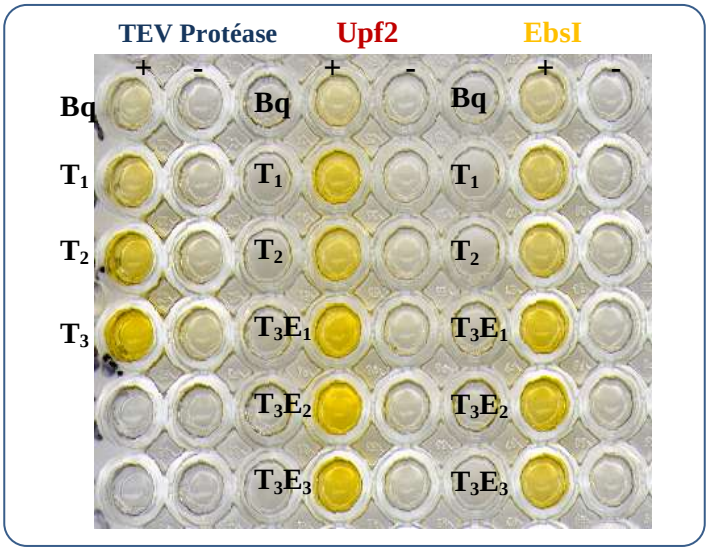


Fig. 3 : Plaque de phage Elisa en *pool* comparant l'enrichissement en phages spécifiques à chaque cible entre les *Output* des différents tours de sélection.

Par comparaison de l'intensité de la coloration jaune nous constatons un net enrichissement de phages spécifiques à la TEV protéase entre la banque naïve, le Tour1, le Tour2 et le Tour3. La spécificité d'interaction des phages avec la cible est montrée dans

l'absence de signal dans les puits de la colonne adjacente qui n'ont pas été recouverts par la protéine cible mais incubés avec le même échantillon de phage. Nous observons un enrichissement en phages spécifiques pour les cibles EbsI et Upf2 entre les différents tours de sélection. Aucune différence de signal n'est visible entre les élutions du Tour3. Ces observations ont été confirmées par la détermination de l'absorbance à 450nm. Les valeurs obtenues sont reportés sur le tableau suivant (Tableau. 2).

	TEV	Upf2	EbsI
Banque	**0,176**	**0,194**	**0,204**
T_1	**0,368**	**0,741**	**0,437**
T_2	**0,608**	**0,467**	**0,399**
T_3E_1	**0,855**	**0,735**	**0,71**
T_3E_2	------	**0,991**	**0,685**
T_3E_3	------	**0,851**	**0,601**

Tableau. 2 : Tableau récapitulatif des absorbances mesurées pour les différents tours de sélection.

Ces mesures montrent une augmentation du signal ce qui revient à l'augmentation du nombre de phages spécifiques à la cible au fil des tours de sélection. On observe aussi une absorbance qui varie différemment entre les trois types d'élutions du tour 3 pour EbsI et Upf2.

Nous avons mis en évidence un enrichissement en clones spécifiques entre les différents tours de sélection. Il est important maintenant d'identifier des clones isolés qui reconnaissent la cible. Ceci va nous permettre par la suite de produire ces variants et de caractériser leurs éventuelles interactions.

3.2. Test de phage Elisa clonal

Le test de phage Elisa clonal est un test qui s'approche du test décrit précédemment mais qui est réalisé à partir de phages provenant de clones isolés (plutôt que de la population comme ci-dessus) prélevés lors de chaque tour de sélection. En effet, des clones issus des différents tours de sélections sont repiqués et les phages correspondants sont produits individuellement. Ces phages clonaux sont mis en contact avec la cible. Ceux qui restent liés à la cible immobilisée, après les lavages, sont révélés par un anticorps anti-M13.

Ce test a été réalisé en plusieurs étapes :

- Culture de clones isolés représentatifs des différents tours de sélection et production des phages correspondants

- Préparation des plaques Elisa
- Test d'interaction et révélation des plaques

a - Préparation des plaques de cultures et production des phages de clones isolés

Pour obtenir des clones isolés, nous avons utilisé les boîtes de comptage : les boîtes de comptage des phages *Input* de la banque naïve et celles du comptage des phages issus des Tour1, Tour2 et les différentes élutions du Tour3.

Les colonies ont été repiquées dans une plaque de culture de 96 puits dans 150 µL de milieu 2YT+Amp+Tet+Glu, pour les différentes cibles, selon le plan présenté sur la Fig. 4.

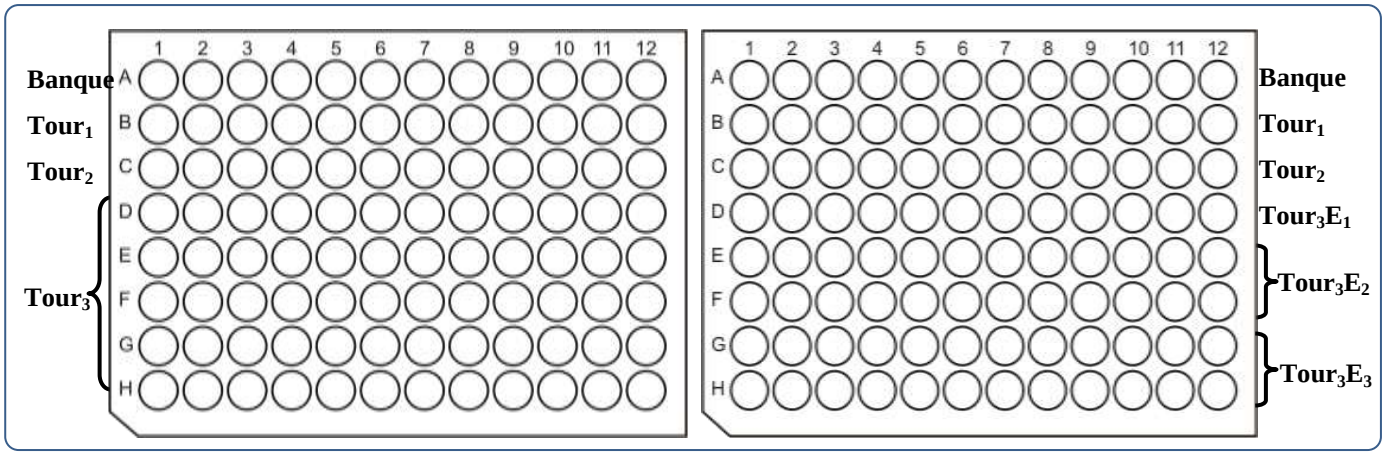


Fig. 4 : Plan de la préparation des plaques de culture des clones issus des différentes étapes de la sélection. Le plan à gauche correspond aux clones de la sélection contre la TEV protéase. Le plan à droite correspond aux deux plaques des clones de la sélection contre Upf2 et EbsI.

Les clonies repiquées, selon ces plans, ont été incubées o.n. à 37°C et 500 rpm. Ainsi nous obtenons la ***« plaque référence »*** *qui a été conservée à -80°C après ajout de 40 µL de milieu 2YT+75% Glycérol par puits.*

Ces plaques ont été répliquées dans d'autres plaques, désignées ***« plaques d'infection »****, en ensemençant 150 µL 2YT+Amp+Tet par puits, avec 10 µL du puits correspondant à la* ***« plaque référence »****. Les plaques d'infection ont été, tout d'abord, incubées 3h à 37°C et 500 rpm. Par la suite, 6µL de phages helper ont été ajoutés à chaque puits et incubés 30min sans agitation puis 30 min à 150 rpm. Ces* ***« plaques d'infection »*** *ont été utilisées par la suite pour ensemencer les* ***« plaques de culture »*** *(100 µL). Ces dernières ont été remplies avec 1.5 mL de 2YT+Amp+Kan. Les phages sont ainsi produits o.n. à 30°C et 150 rpm. Puis ils sont récupérés par centrifugation des plaques 1 h à 2000 g. Toutes ces étapes sont résumées dans la Fig. 5.*

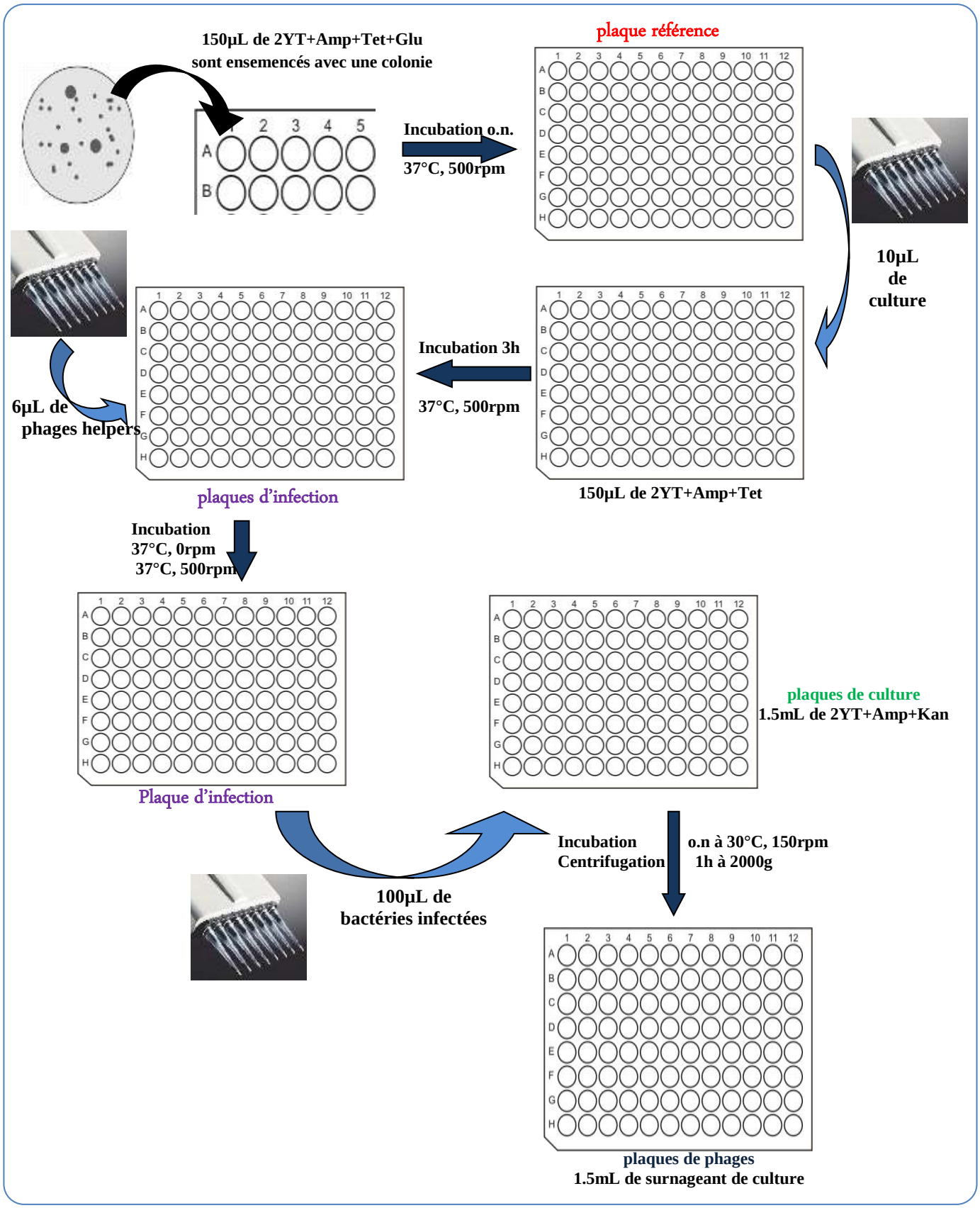


Fig. 5 : Schéma général du protocole suivi pour la production des phages clonaux utilisés pour le test *phage Elisa clonal*.

b - Préparation des plaques Elisa

6 plaques Elisa ont été préparées, pour effectuer les tests d'interaction des phages clonaux avec les cibles.

Trois plaques ont été incubées o.n. à 4°C et 300 rpm avec 150 µL/puits d'une solution de protéine cible à 20 µg/mL. Les 3 autres plaques ont été utilisées comme plaques témoins négatifs : 2 plaques ont été incubées avec du tampon PBS témoin négatif pour TEV protéase et EbsI alors que la troisième a été incubée au tampon acétate. Ces plaques ont été bloquées par la suite avec 250 µL de TBST BSA 4% pendant 3 h à 4°C puis elles ont été lavées 3 fois avec 250 µL TBST.

c - Test d'interaction et révélation des plaques Elisa

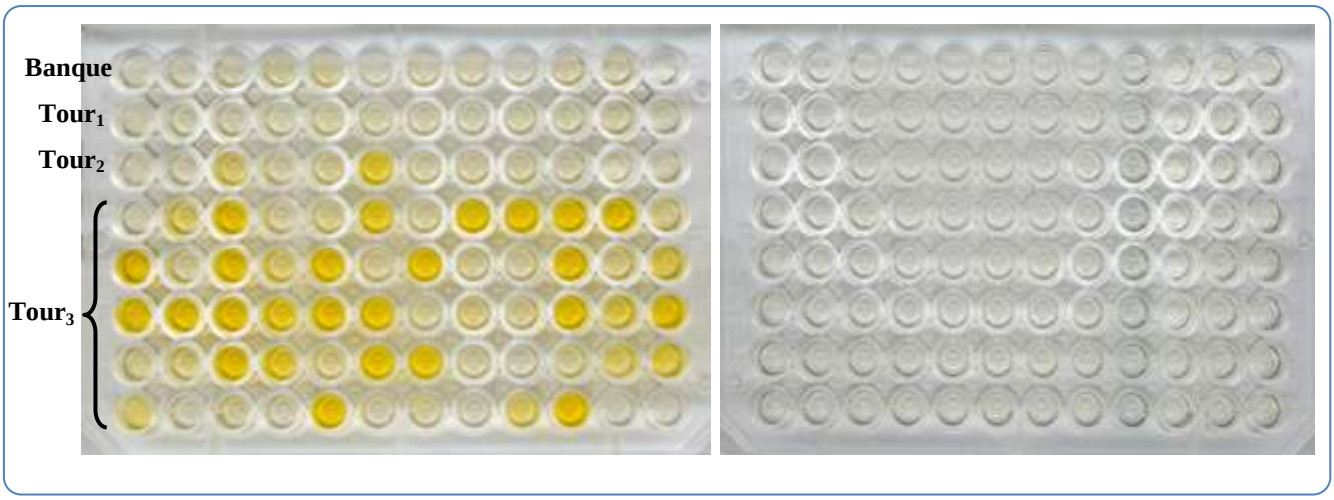
Les plaques Elisa ainsi préparées ont été par la suite incubées, avec 100 µL de phages des surnageants de cultures. Après une incubation de 2 h à 500 rpm et 4°C, 4 lavages avec une solution TBST ont été réalisés pour éliminer les phages non spécifiques. Les phages qui restent fixés à la cible ont été révélés par incubation 1 h à 25°C, avec 100 µL d'une solution de TBST contenant l'anticorps anti-M13 couplé à la peroxydase à la dilution 1/5000. Après 4 lavages au TBST, 100 µL de substrat soluble ont été ajoutés dans chaque puits des plaques Elisa. Une coloration jaune apparaît suite à l'hydrolyse du substrat en présence de la peroxydase liée à l'anticorps reconnaissant un phage exposant à sa surface une αRep spécifique à la cible immobilisée. La réaction a été stoppée par ajout de 100 µL d'acide HCl 1 N.

Les résultats obtenus sont résumées dans la partie suivante.

- *Test phage Elisa clonal de la TEV protéase*

Nous avons choisi de tester des clones des différents tours de sélection ainsi que de la banque « naïve ». Ce test a pour but d'identifier des clones isolés qui reconnaissent la TEV protéase et de vérifier si la proportion de clones positifs augmente lors des cycles de sélection. Les clones positifs sont mis en évidence par interaction entre la cible immobilisée et l'*αRep* exposée sur phage.

Les plaques Elisa de la Fig. 7 montrent clairement l'enrichissement en clones spécifiques à la TEV protéase entre la banque naïve et les autres tours de sélection : Aucun clone spécifique n'a été identifié à partir de banque naïve et du Tour1, 2 clones ont été identifiés à partir du Tour2 dont un avec un signal faible alors qu'au Tour3 on peut compter 32/60 clones nettement positifs. Ces plaques confirment l'absence de signal sur la plaque témoin et donc la spécificité de la reconnaissance de ces clones isolés pour la protéine cible, plutôt que pour le support lui-même.



Plaque couverte avec la TEV protéase **Plaque sans TEV protéase (témoin négatif)**

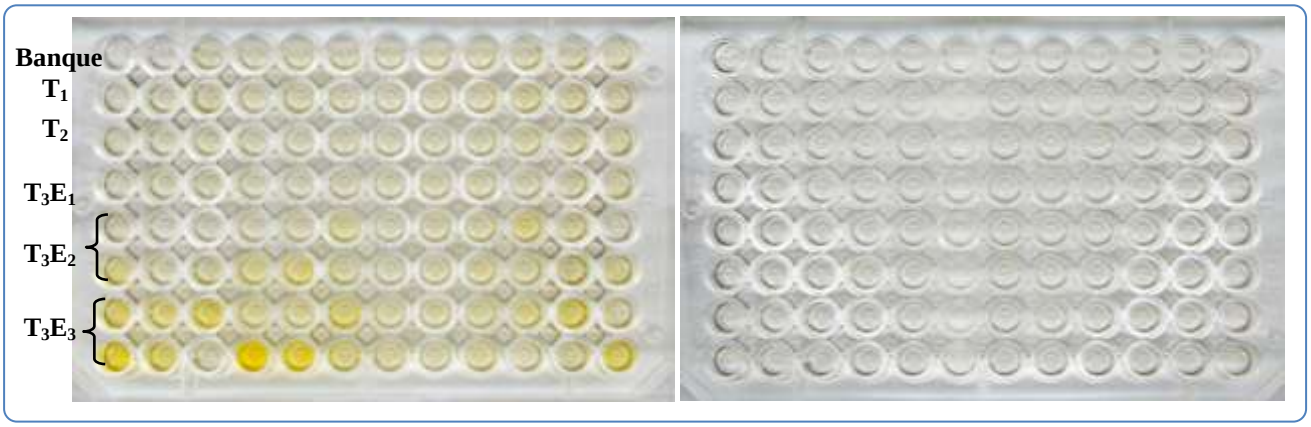
Fig. 7 : Les plaques Elisa révélées montrant les puits correspondants aux phages exposant les variants qui reconnaissent la cible.

14 clones, issus des différents tours de sélection, ont été retenus pour une étude plus détaillée des interactions avec la TEV protéase (Fig. 8).

Banque	0,105	0,072	0,070	0,159	0,150	0,049	0,063	0,088	0,022	0,107	0,102	-0,001
T_1	0,172	0,067	0,055	0,073	0,123	0,077	0,073	0,062	0,049	0,063	0,099	0,064
T_2	0,099	0,053	0,421	0,074	0,131	0,838	0,076	0,084	0,074	0,055	0,059	0,062
T_3	0,112	0,448	1,443	0,091	0,183	0,926	0,046	1,204	1,565	1,212	1,510	0,235
	1,921	0,091	1,535	0,300	1,675	0,070	1,571	0,074	0,043	1,556	0,169	0,616
	1,670	1,369	1,444	0,731	1,807	1,609	0,054	0,202	0,108	1,131	0,512	1,707
	0,215	0,227	1,534	0,734	0,139	1,330	1,179	0,065	0,011	0,118	0,416	0,723
	0,588	0,023	0,136	0,034	1,478	0,042	0,106	0,014	0,420	1,710	0,013	0,038

Fig. 8 : Ce tableau représente la plaque Elisa et regroupe les absorbances à 450nm mesurées dans les différents puits. Les clones qui seront retenus pour la caractérisation des interactions sont entourés par un cercle.

❖ *Test phage Elisa clonal de la cible Upf2*



Plaque couverte avec l'Upf2 **Plaque sans Upf2 (témoin négatif)**

Fig. 9: Les plaques Elisa révélées montrant les puits correspondants aux phages qui reconnaissent la cible.

Ce test de phage Elisa clonal montre que très peu de clones reconnaissent la cible (13/96). Ils sont essentiellement issus du tour 3 et en particulier de l'élution non-spécifique (l'élution acide) (Fig. 9).

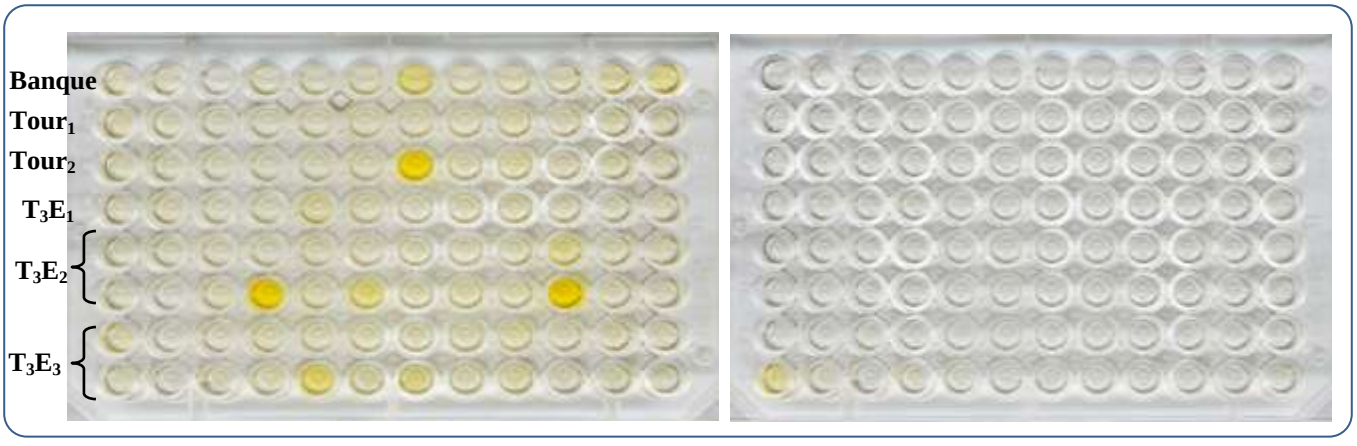
Les absorbances à 450 nm des différents puits ont été mesurées et nous n'avons retenu que 8 clones qui seront utilisés pour la caractérisation des interactions (Fig. 10) dont 2 sont issus de l'élution faite par compétition avec la cible libre (Elution2) et 6 proviennent de l'élution acide.

Banque	0,030	-0,002	0,044	0,055	0,050	0,044	0,047	0,054	0,052	0,067	0,059	0,074
T_1	0,023	0,039	0,056	0,037	0,052	0,021	0,036	0,045	0,044	0,093	0,073	0,077
T_2	0,045	0,033	0,041	0,036	0,031	0,043	0,028	0,037	0,046	0,081	0,064	0,085
T_3E_1	0,044	0,042	0,038	0,042	0,023	0,041	0,036	0,056	0,049	0,049	0,052	0,077
T_3E_2	0,041	0,003	0,006	0,030	0,049	**0,122**	0,032	0,069	0,070	**0,181**	0,091	0,000
	0,197	0,038	0,063	**0,130**	**0,176**	0,057	0,030	0,059	0,093	0,085	**0,110**	**0,127**
T_3E_3	**0,289**	**0,176**	**0,288**	0,098	0,091	**0,211**	0,080	0,075	0,087	**0,147**	**0,262**	0,098
	0,536	**0,233**	0,011	**0,890**	**0,456**	**0,175**	**0,103**	0,069	**0,112**	0,062	0,084	**0,373**

Fig. 10: Ce tableau représente la plaque Elisa et regroupe les absorbances à 450nm mesurées dans les différents puits. Les clones qui seront retenus pour la caractérisation des interactions sont entourés par un cercle.

❖ *Test phages Elisa clonal de la cible EbsI*

Comme pour les deux cibles précédentes, un test phage Elisa clonal (Fig. 11) a été réalisé pour EbsI dans le but d'identifier des clones isolés reconnaissant spécifiquement la cible EbsI. Ces clones seront par la suite produits pour faire des tests d'interactions permettant d'évaluer la constante de dissociation du complexe (biacore, ITC…..).



Plaque immobilisée avec EbsI **Plaque sans EbsI (témoin négatif)**

Fig. 11: Les plaques Elisa révélées montrant les puits correspondants aux clones exposés sur phages reconnaissant la cible.

Très peu de clones sont identifiés pour EbsI. Mais ce qui est différent par rapport aux tests précédents est l'apparition de deux clones spécifiques très tôt à partir du second tour de sélection et même de la banque. Nous prendrons en compte ces clones dans le but d'évaluer l'évolution de la population au cours des différents tours de la sélection.

L'observation de la coloration jaune, témoignant de la présence de phages qui exposent une *αRep* reconnaissant la cible, a été par la suite confirmée par la mesure de l'absorbance à 450nm. Ceci nous a permis de choisir 6 clones pour l'étude et la caractérisation des interactions Fig. 12.

Banque	0,080	0,044	0,009	0,059	0,056	0,048	**0,275**	0,059	0,068	0,050	0,108	**0,202**
Tour$_1$	0,085	0,052	0,031	0,031	0,041	0,051	0,051	0,043	0,069	0,052	0,053	0,054
Tour$_2$	0,068	0,035	0,036	0,032	0,037	0,046	0,872	0,041	0,089	0,042	0,071	0,064
T$_3$E$_1$	0,050	0,056	0,023	0,037	0,136	0,062	0,023	0,040	0,054	0,035	0,031	0,045
T$_3$E$_2$	0,054	0,039	0,037	0,022	0,062	0,046	0,057	0,048	0,049	0,180	0,049	0,064
	0,052	-0,003	0,050	0,913	0,051	0,265	0,051	0,093	0,039	1,059	0,045	0,050
T$_3$E$_3$	0,123	0,010	0,050	0,063	0,052	0,054	0,057	0,061	0,083	0,029	0,059	0,040
	-0,143	-0,001	0,048	0,043	0,328	0,081	0,185	0,057	0,098	0,029	0,052	0,032

Fig. 12: Ce tableau représente la plaque Elisa et regroupe les absorbances à 450nm mesurées dans les différents puits. Les clones qui seront retenus pour la caractérisation des interactions sont entourés par un cercle.

3.3. Récapitulatif des résultats de la sélection

Après les sélections, nous avons pu identifier 14 interacteurs potentiels de la TEV protéase, 8 pour Upf2 et 6 pour EbsI. Les clones retenus seront par la suite caractérisés : leurs plasmides seront extraits et séquencés dans le but de vérifier s'il n'existe pas de séquences redondantes. Nous testerons par la suite l'expression et la solubilité de ces variants dans le but de tester la spécificité de l'interaction cible-*αRep* libre.

Dans le but de vérifier la diversité des clones identifiés, les plasmides correspondants ont été tout d'abord extraits puis digérés NdeI-HindIII.

Ceci nous a permis de déterminer le nombre de motifs insérés par protéine et la distribution de la taille des variants s'est révélée très variable.

En effet, pour :

- TEV protéase : Les 14 interacteurs potentiels contiennent entre 2 et 9 motifs insérés avec une majorité de clones avec 5 motifs.

- Upf2 : Les 8 interacteurs potentiels contiennent entre 3 et 8 motifs insérés avec une majorité de variants à 7 motifs.
- **EbsI** : 6 interacteurs potentiels contenant entre 2 et 7 motifs dont 2 ont 6 motifs insérés.

Les séquences de ces différents mutants ont été déterminées. Par alignement de séquence, nous avons vérifié que les clones, incorporant le même nombre de motifs, ont des séquences différentes. Nous montrons sur la Figure. 13, un exemple d'alignement de 2 interacteurs potentiels d'EbsI, identifiés lors du test phage Elisa clonal.

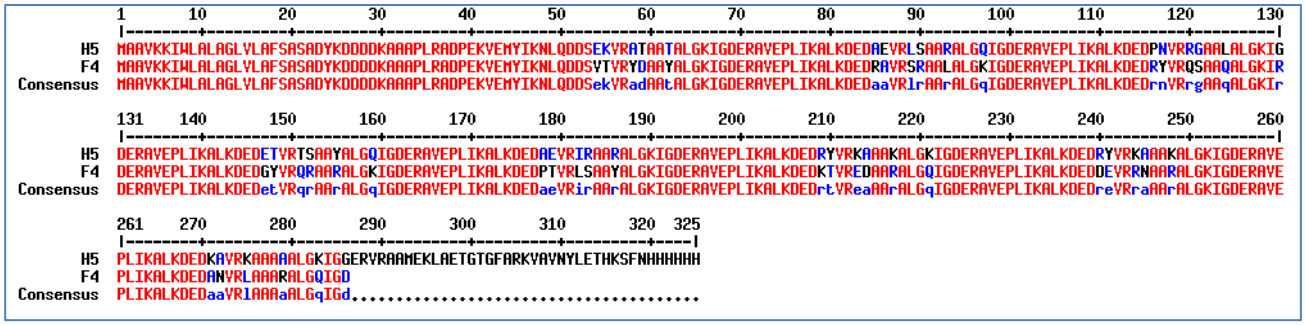


Fig. 13 : Alignement de 2 interacteurs de la protéine EbsI, contenant 7 motifs.

Cet alignement montre clairement que ces deux mutants, bien qu'ils contiennent le même nombre de motifs, ont des séquences différentes. Ceci a été vérifié pour tous les autres interacteurs des différentes cibles.

Nous nous sommes fixés comme objectif, par la suite de trouver des tests qui permettant de cribler ces variants pour ne retenir que ceux qui ont les meilleures affinités pour les étapes de caractérisation ultérieures.

4. Criblage des interacteurs potentiels de la TEV protéase, Upf2 et EbsI identifiés par les tests de phage Elisa clonal

Ce criblage secondaire a été conçu pour comporter deux étapes : une première étape est un Test Elisa pour confirmer l'interaction cible-*αRep* libre, c'est-à-dire non exposée sur phage, et un deuxième Test Elisa réalisé avec différents temps de lavage, dans le but d'identifier les protéines ayant une dissociation lente.

4.1. Test de reconnaissance cible-*αRep*

Ce test de reconnaissance cible-*αRep* libre est un test Elisa qui est basé sur la reconnaissance de l'interacteur *αRep* avec la cible immobilisée sur la plaque Elisa. Une fois

l'*αRep* fixée à la cible, elle est révélée par le biais de son étiquette *Flag*. Cette étiquette est reconnue par un anticorps anti-*Flag-tag*.

Trois étapes sont essentielles pour la réalisation de ce test :

- Préparation de la plaque Elisa
- Préparation des interacteurs
- Test d'interaction et révélation

a - Préparation de la plaque Elisa

Pour ce test, une plaque Elisa a été tout d'abord couverte par la cible selon le plan présenté ci-dessous.

En effet, la plaque a été divisée en trois parties (Fig. 14). Les colonnes (+) correspondent aux puits sur lesquels la cible a été immobilisée. Ces puits ont été incubés o.n. à 4°C avec 100 µL d'une solution de la cible à 20 µg/mL. Alors que les colonnes (-) correspondent aux puits témoins négatifs qui ont été incubés, en parallèle, avec 100 µL d'une solution de tampon PBS.

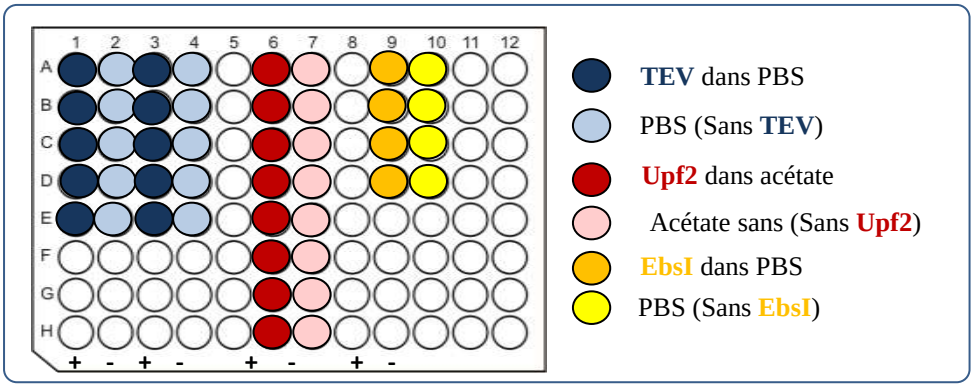


Fig. 14 : Plan de préparation de la plaque Elisa avec les puits immobilisés avec les cibles et ceux utilisés comme témoin négatifs.

Après l'immobilisation des cibles, les puits de la plaque ont été lavés 3 fois avec TBST puis bloqués par incubation 2 h à 4°C et 500 rpm avec 200 µL d'une solution de TBST BSA 3%.

b - Préparation des interacteurs

Plusieurs méthodes de simplification du test ont été essayées. Le test le plus opérationnel utilisé est basé sur l'expression par le promoteur T7 présent sur le phagemide dans la souche *Rosetta Blue*. Les protéines sont exprimées de façon isolée, sans fusion avec la protéine III, puis secrétées. On utilise pour cela une souche exprimant la T7 polymérase. Les protéines exprimées après induction sont testées directement à partir de lysats bactériens obtenus par traitement des culots cellulaires avec un mélange de lyse basé sur des détergents.

En effet, les plasmides des différents interacteurs ont été utilisés pour transformer des bactéries Rosetta Blue. Les bactéries transformées ont été utilisées pour ensemencer une préculture de 10 mL de 2YT+Amp+Glu. Les précultures ont été utilisées par la suite pour préparer des cultures à DOi=0.1. Elles ont été incubées à 37°C, 220 rpm. A DO=0.6, l'expression des interacteurs a été induite par ajout d'IPTG. Après 4 h d'induction, les bactéries ont été collectées par centrifugation. Elles sont lysées par ajout de 500 µL de tampon de lyse B-PERII et 3 cycles de congélation-décongélation. Les interacteurs ont été récupérés à partir des fractions solubles après centrifugation (20 min, 14000 rpm). Ces fractions solubles ont été utilisées pour le test Elisa.

c - Test d'interaction et révélation

Dans le but de vérifier les interactions cible-αRep libre, 200 µL des fractions solubles de chaque interacteur ont été prélevés dont 100 µL sont incubés dans les puits immobilisés avec la cible et 100 µL sont incubés dans les puits (-). Après 1 h 30 min d'incubation à 4°C et 4 lavages au TBST, les αReps qui interagissent avec la cible ont été révélées par incubation, 1 h à 4°C, avec 100 µL d'une solution de TBST contenant l'anticorps anti-Flag-tag couplé à la peroxydase dilué au 1/20000. Après 4 lavages au TBST, 100 µL de substrat soluble ont été ajoutés. La coloration bleue avait été neutralisée par ajout de 100 µL de HCl 1 N (Fig. 15).

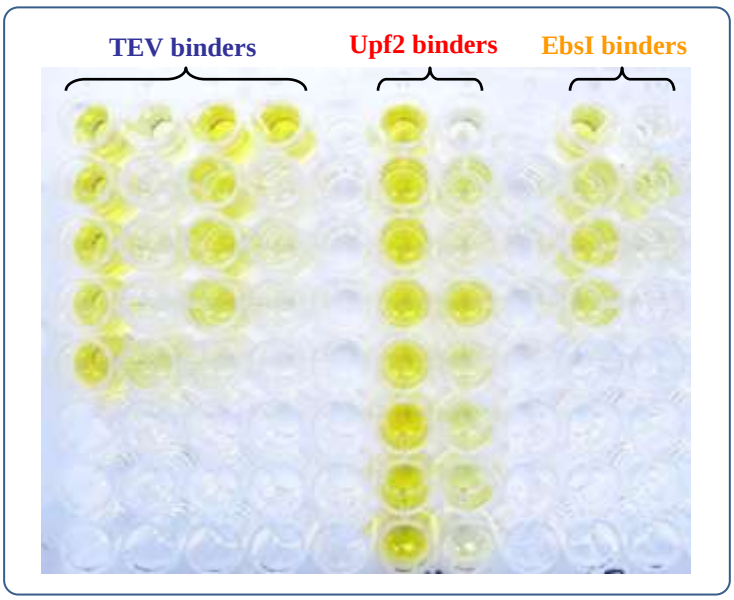
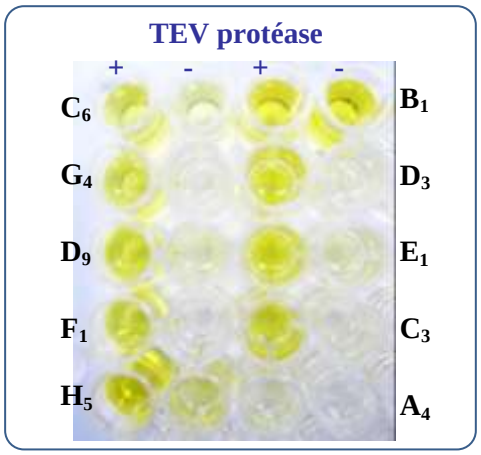


Fig. 15 : Plaque Elisa révélée montrant l'interaction entre les différents interacteurs retenus et les cibles correspondante. Les puits recouverts des protéines cibles sont disposés en colonne et alternent avec des colonnes recouvertes de BSA ce qui permet d'identifier des clones non spécifiques.

Cette plaque Elisa nous a permis d'étudier la spécificité de l'interaction des cibles avec les différents interacteurs. Nous allons passer en revue, dans la partie suivante les interactions observées chaque cible à part.

❖ *Les interacteurs de la TEV protéase*

La figure. 16 montre les signaux obtenus avec les différents interacteurs de la TEV protéase avec la mesure de l'absorbance à 450nm.

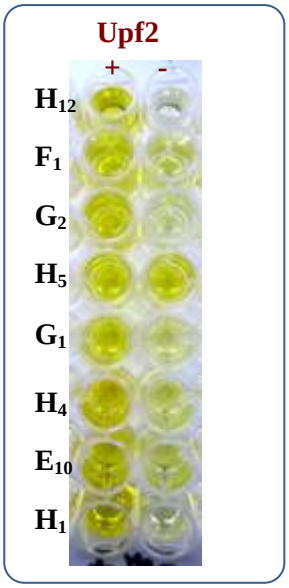


C_6	**0,161**	**-0,027**	**B_1**
G_4	**0,279**	**0,323**	**D_3**
D_9	**0,275**	**0,288**	**E_1**
F_1	**0,295**	**0,277**	**C_3**
H_5	**0,330**	**0,046**	**A_4**

Fig. 16 : Plaque Elisa de l'interaction TEV avec 10 interacteurs *αRep* et un tableau récapitulatif (à droite) des absorbances à 450nm mesurées.

Ces résultats confirment que l'interacteur A_4 n'interagit pas ou peu avec la TEV protéase : faible signal en contact de la TEV. L'interacteur B_1 semble reconnaître la BSA aussi efficacement que la TEV. Ces deux interacteurs ont été prélevés respectivement à partir de la banque naïve et à l'issue du 1er tour de sélection ce qui suggère que les clones vraiment spécifiques n'apparaissent qu'après deux tours de sélection. Les 8 interacteurs restant présentent des signaux positifs d'intensités différentes. Ils seront retenus pour le Test Elisa à différents temps de lavages.

❖ *Les interacteurs d'Upf2*

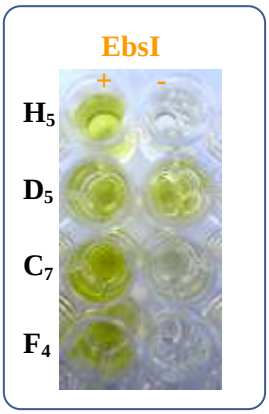


H_{12}	**0,379**
F_1	**0,281**
G_2	**0,395**
H_5	**0,044**
G_1	**0,307**
H_4	**0,539**
E_{10}	**0,429**
H_1	**0,542**

Fig. 17 : Plaque Elisa de l'interaction Upf2 avec 8 interacteurs *αRep* et un tableau récapitulatif (à droite) des absorbances à 450nm mesurées.

Les résultats de la Fig. 17 nous montrent que l'interacteur H_5 n'est pas spécifique à Upf2: en effet il donne une intensité de signal similaire en présence ou en absence d'Upf2. Il n'est donc pas retenu pour le test Elisa avec différents temps de lavage. Les 7 interacteurs restant seront alors utilisés pour les étapes suivantes de caractérisation des interactions.

❖ *Les interacteurs d'EbsI*



H_5	**0,221**
D_5	**0,005**
C_7	**0,254**
F_4	**0,191**

Fig. 18 : Plaque Elisa de l'interaction EbsI avec 4 interacteurs *αRep* et un tableau récapitulatif (à droite) des absorbances à 450 nm mesurées.

Pour cette cible (EbsI), nous retenons trois interacteurs ayant de faibles signaux. L'*αRep* D_5 n'est pas retenue car elle n'est pas spécifique à la cible.

Ce test d'interaction cible-*αRep* libre nous a permis effectivement de faire un premier crible des différents interacteurs identifiés par phage Elisa clonal en excluant ceux qui ne reconnaissent pas de façon spécifique le cible correspondante.

Toute fois les différences d'intensité des signaux ne peut pas être considérée comme un indicateur de l'affinité interacteur/cible. En effet les protéines testées sont prélevées directement des fractions solubles des bactéries lysées : l'intensité des signaux peut varier selon l'affinité du binders à la cible mais aussi la quantité de protéine exprimée qui diffère d'une culture à une autre. Ainsi nous avons pensé à réaliser un Test Elisa avec différent temps de lavage qui nous permettra de comparer les interacteurs entre eux et avoir une estimation qualitative de l'affinité.

4.2. Test d'estimation qualitative de l'affinité

Ce test vise le criblage des interacteurs identifiés par phage Elisa clonal par une estimation qualitative de l'affinité donnée par la variation de l'intensité du signal Elisa en fonction du temps de lavage. Ce test est réalisé en trois étapes :

- Préparation des plaques Elisa.
- Préparation des interacteurs.
- Test d'interaction et révélation

a - Préparation des plaques Elisa

Les 2 plaques ont été préparées, comme décrit au test précèdent.

Le plan des plaques est présenté dans la Fig. 19 :

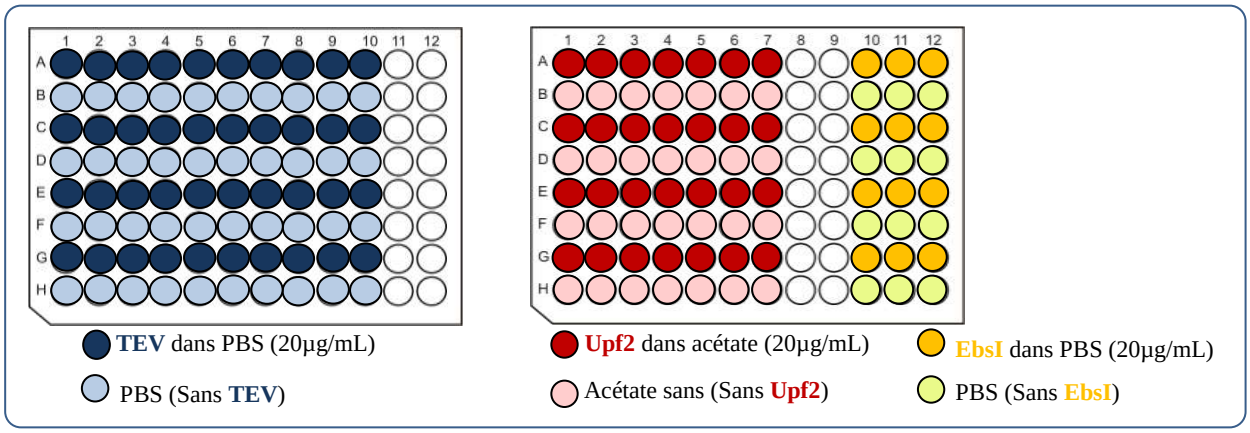


Fig. 19: Plan selon lequel les plaques Elisa ont été incubées avec les cibles.

b - Préparation des interacteurs

Les échantillons des interacteurs ont été préparés comme décrit dans le (a) pour le Test Elisa. En effet, les échantillons d'interacteurs sont prélevés directement de la fraction soluble (FS) des lysats bactériens.

c - Test Elisa et révélation

Le but de ce test est de visualiser et quantifier dans la mesure du possible l'effet du lavage sur l'interaction de l'interacteur avec la cible. En effet, les différents puits (avec ou sans cible) sont incubés dans les mêmes conditions avec l'interacteur testé. La différence entre les puits réside dans le temps de lavage : les interacteurs les plus affins doivent alors rester fixés à la cible même après un temps de lavage assez long, alors que ceux qui ne sont pas ou peu affins sont libérés et éliminés pour des temps de lavages plus courts. Ce test donne une évaluation relative du paramètre k_{off} de l'interaction. Les affinités élevées étant

caractérisées par des constantes individuelles de dissociation lentes. L'estimation du k_{off} étant liée à la décroissance temporelle du signal et non à son amplitude initiale elle ne dépend pas de la quantité protéine exprimée.

Sur le plan pratique, les puits sont incubés 1h en présence de la cible avec des intervalles temporaires. Chaque colonne de la plaque correspond à un interacteur et chaque paire de lignes correspond à un temps de lavage. Nous avons procédé de la façon suivante :

- *Incubation avec 100 µL de la FS d'interacteur de T_0 à T_0+1 h*
- *Lavage avec 200 µL de TBST de T_0 + 1h à T_0+2 h 30 min*

} **Lignes A et B**

- *Incubation avec 100 µL de FS d'interacteur de T_0+30min à T_0+1 h 30min*
- *Lavage avec 200 µL de TBST de T_0+1 h 30 min à T_0+2 h 30 min*

} **Lignes C et D**

- *Incubation avec 100 µL de FS d'interacteur de T_0+1 h à T_0+2 h*
- *Lavage avec 200 µL de TBST de T_0+2 h à T_0+2 h 30 min*

} **Lignes E et F**

- *Incubation avec 100 µL de FS d'interacteur de T_0+1 h 15 min à T_0+2 h 15 min*
- *Lavage avec 200 µL de TBST de T_0+2 h 15min à T_0+2 h 30 min*

} **Lignes G et H**

Selon ce plan, les lignes A et B correspondent à une durée de lavage de 1h30, les lignes C et D à un temps de lavage de 1h, les lignes E et F ont été lavées 30' alors que les lignes G et H n'ont été lavés que pendant 15' de lavage seulement. Par la suite, la révélation des protéines fixées à la cible est réalisée comme décrit dans la partie (a) par un anticorps anti-*His-tag* couplé à la peroxydase.

Nous passons maintenant à la présentation des résultats obtenus : les plaques Elisa à différents temps de lavages et interprétation des signaux pour chaque cible.

❖ *Les interacteurs de la TEV protéase*

Ce test Elisa à différents temps de lavage (Fig. 20) confirme les conclusions émises suite au test Elisa précédent : le clone A_4 est peu ou pas affin à la TEV protéase puisqu'il perd la totalité de son signal au bout de 15min de lavage. Nous vérifions aussi que le clone B_1 reconnait la BSA et non pas la cible. Nous observons aussi que le lavage n'a pas d'effet sur cette interaction.

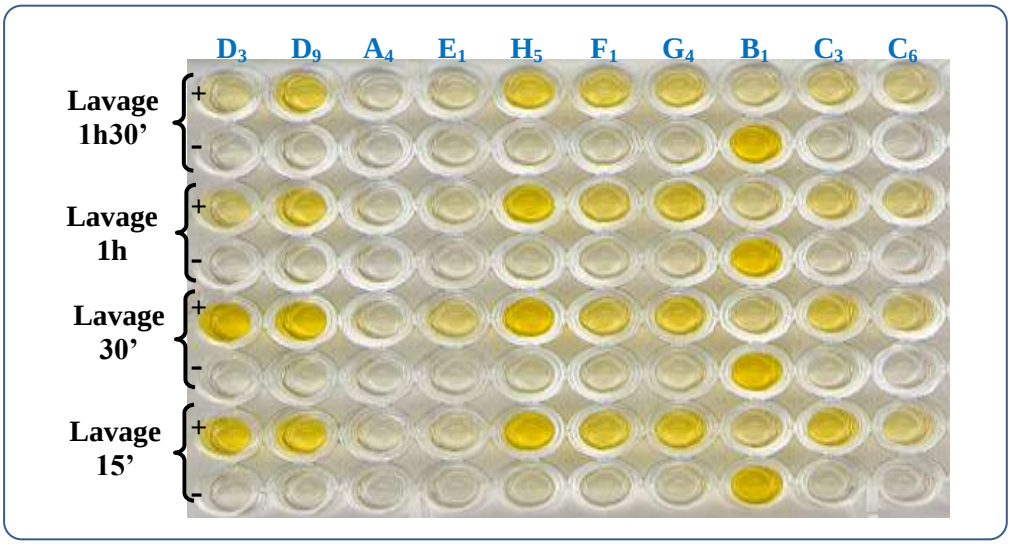


Fig. 20 : Plaque révélée du test Elisa avec les différents temps de lavages pour les interacteurs de la TEV protéase.

Pour les autres interacteurs, nous constatons des comportements variables : Le clone D_3 par exemple voit son signal décroitre significativement au bout d'une heure de lavage alors que le clone D_9 garde un signal important même après 1h30' de lavage. Ces comportements nous permettent de comparer l'affinité des clones. Il semble que le clone D_9, H_5, F_1, sont les clones les plus affins suivis par les clones D_3, G_4, C_3, C_6, et E_1. Ces observations ont été confirmées par la mesure des absorbances à 450 nm (Fig. 21).

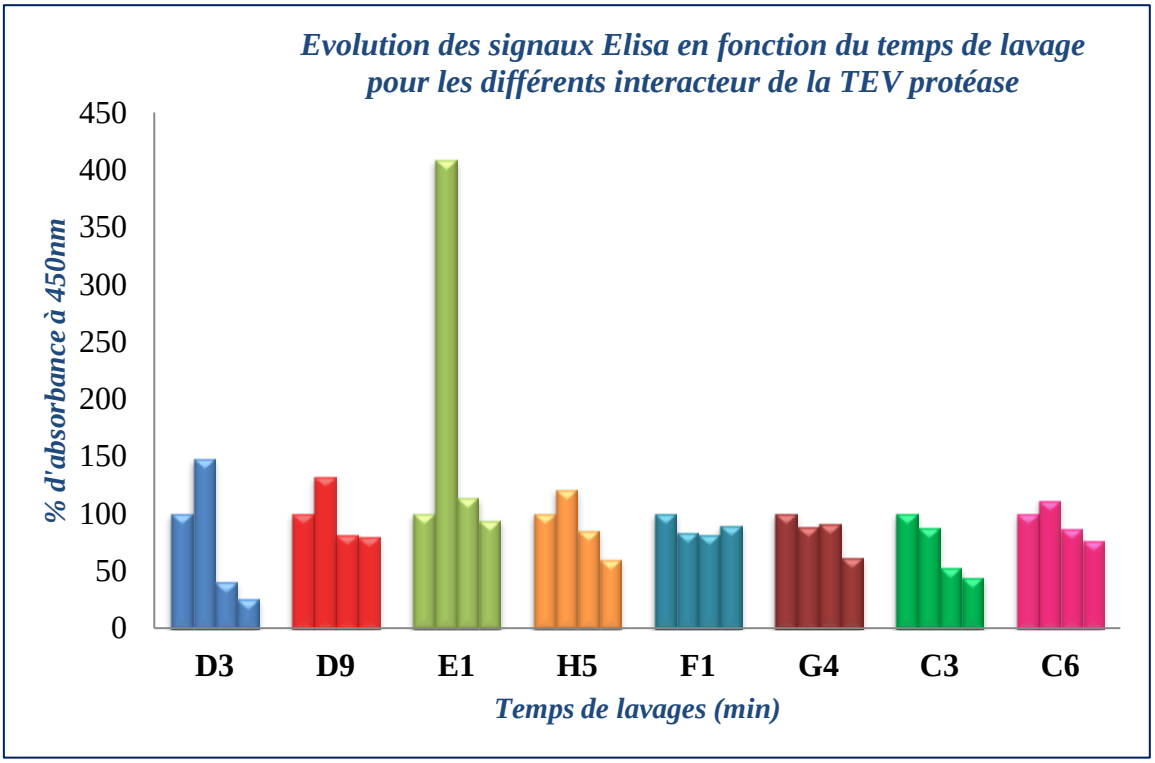


Fig. 21 : Evolution de l'absorbance à 450nm en fonction du temps de lavage pour les différents interacteurs spécifiques à la TEV protéase.

Ce graphe représente l'évolution des absorbances à 450 nm et par conséquent la quantité d'*αRep* liées à la cible en fonction du temps de lavage. Nous avons pris en compte, pour tracer ce graphe, un % d'absorbance qui a été calculé en désignant le signal à 15min comme signal total (100 %).

Tout d'abord, nous constatons, pour les clones D3, D9 E1, H5et C6, que l'intensité du signal après 30 min de lavage est plus importante que celle mesurée après 15 min de lavage. Les interacteurs *αRep* sont prélevés directement du lysat bactérien, l'échantillon contient alors en plus des *αRep* surexprimées des protéines solubles des bactéries, de l'ARN, de l'ADN et tous les métabolites exprimés dans les bactéries, ces contaminants pourraient interférer lors de la détection *αRep*-cible. Lorsque le temps et/ou le nombre de lavages augmente ces contaminants pourraient être éliminés ce qui pourrait diminuer ces interférences et par la suite donner un signal Elisa plus important. Nous avons aussi observé pour d'autres tests Elisa *αRep*-cible (résultats non publiés) que les échantillons de lysat bactériens préalablement dilués donnent des signaux Elisa qui sont plus importants que les échantillons non dilués. Cet effet, n'est pourtant par observé pour les 3 clones restant ce qui est peut être dû à d'autres variations survenues lors des cultures bactériennes.

Nous observons aussi qu'une disparition du signal au fil du temps de lavage est différente d'un clone à un autre ce qui reflète une potentielle différence d'affinité entre les clones. En effet, un clone affin sera détaché de la cible à un temps de lavage plus important que celui qui est moins affin.

Parmi les 8 *αRep* choisies, le clone F1 semble être l'interacteur le plus affin : il a été identifié parmi les clones du Tour3 et ayant le signal Elisa le plus important en phage Elisa clonal. Le clone D3 a été choisi parmi les clones du Tour3 ayant un signal en phage Elisa important mais il semble être le moins affin. Pour les clones C3 et C6, qui ont été choisis du Tour2, nous pouvons confirmer que le C6 est plus affin que le C3 comme observé pour les signaux en phage Elisa clonal. Pour les clones restants, nous pouvons estimer un classement selon l'affinité : H5<G4<E1 et D9.

Il faut bien noter que ces résultats restent qualitatifs et destinés à trier les clones les plus affins. Ces résultats doivent être précisés par une mesure de la constante de dissociation par Biacore ou ITC.

❖ *Les interacteurs d'Upf2*

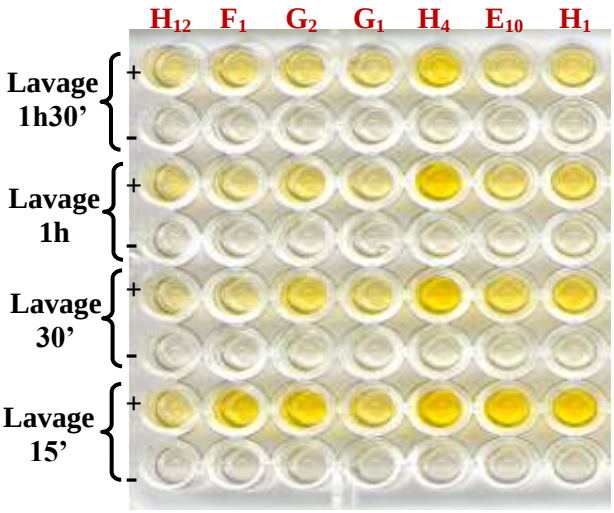


Fig. 22 : Plaque révélée du Test d'interaction d'Upf2 avec les 7 clones d'*αRep* retenus avec différents temps de lavages.

A partir de l'extinction des signaux Elisa en fonction du temps de lavage observée sur la plaque (Fig. 22), nous constatons que le clone H_{12} est pas ou peu affin à Upf2 suivi par les clones G_1, F_1 puis dans le même rang les clones G_2, E_{10}, et H_1. Le clone H_4 semble être le clone le plus affin pour Upf2. Nous avons mesuré aussi les absorbances à 450nm et nous avons tracé les courbes d'évolution de l'absorbance en fonction de la durée du lavage (Fig. 23).

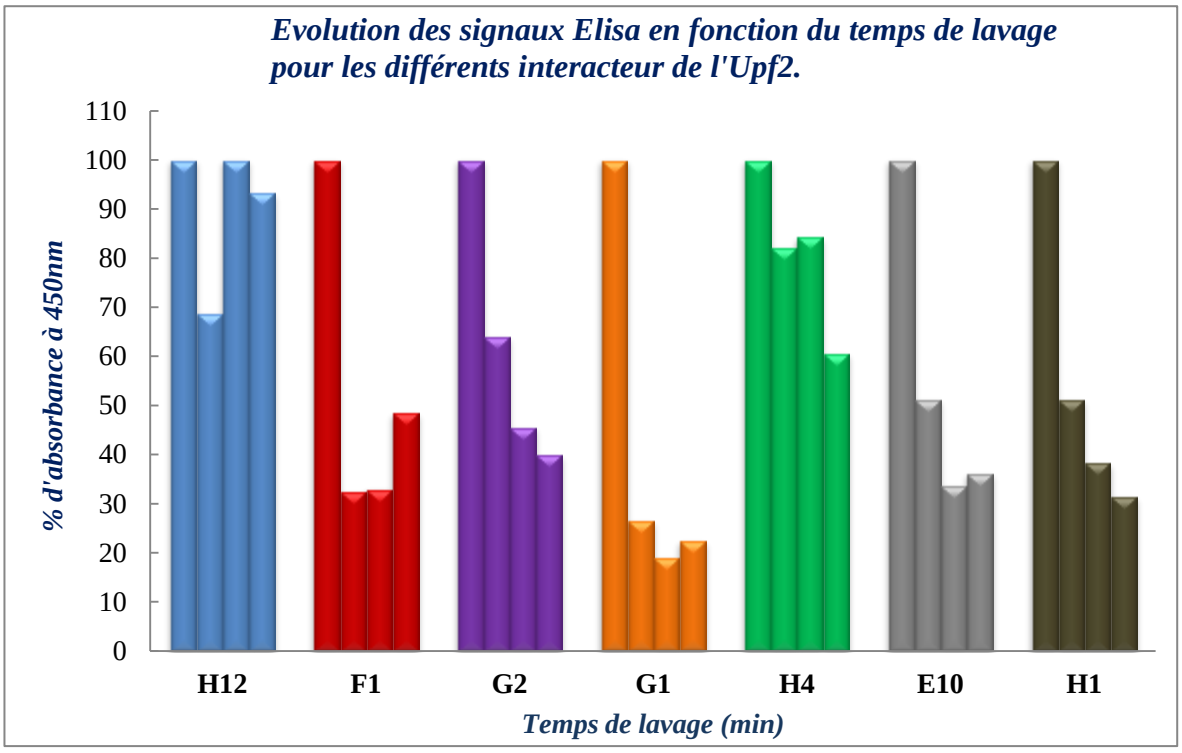


Fig. 23: Evolution de l'absorbance à 450nm en fonction du temps de lavage pour les différents interacteurs spécifique à l'Upf2.

Ce graphe nous permet de confirmer que le clone H4 est le plus affin pour Upf2 ce qui correspond au clone ayant le signal le plus important en phage Elisa clonal. Le clone G2 semble être moins afin que H4. Les clones G2, E10 et H1 semblent avoir des comportements similaires et par la suite des affinités qui seraient proches. Comme pour la TEV protéase, ces résultats restent des résultats comparatifs qui doivent être confirmés par détermination de la constante de dissociation K_D correspondante à chaque interacteur.

❖ *Les interacteurs d'EbsI*

Nous avons procédé de la même manière pour les 3 interacteurs retenus pour EbsI (Fig. 24).

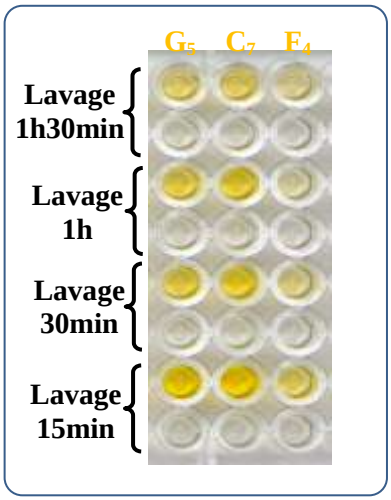


Fig. 24 : Plaque révélée du test d'interaction d'EbsI avec les 3 clones d'*αRep* retenus avec différents temps de lavage.

Pour EbsI, le clone F_4 semble être l'interacteur le moins affin puis qu'il perd la quasi-totalité de son signal à 15min de lavage. La comparaison entre les 2 autres clones nécessite une analyse de l'évolution de l'absorbance à 450nm en fonction des durées de lavage (Fig. 25). Ce graphe nous confirme les résultats décrits au préalable et en revanche, ils ne permettent pas de donner des indications supplémentaires concernant l'affinité des clones C_7 et H_5.

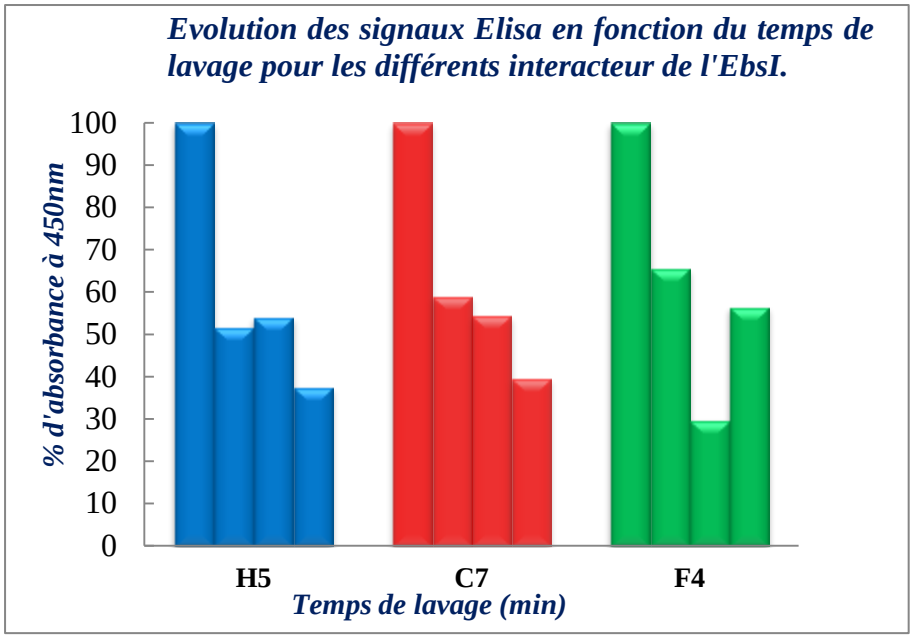


Fig. 25 : Evolution de l'absorbance à 450nm en fonction du temps de lavage pour les différents interacteurs spécifiques à l'EbsI.

La mesure de la constante de dissociation pour les différentes interactions identifiées suppose de disposer d'une quantité plus importante de protéine purifiée. Nous avons alors procédé tout d'abord au sous-clonage des différentes *αRep* dans le plasmide d'expression *pQE*, conduisant à une expression cytoplasmique plus efficace. Les protéines ont été par la suite produites puis purifiées pour pouvoir mesurer les constantes de dissociation des complexes par ITC et Biacore.

Ma propre contribution à ce travail expérimental s'est arrêtée à ce stade, faute de temps….Par ailleurs d'autres sélections ont été réalisées au laboratoire à partir de la banque dont la construction est décrite dans cette thèse. La dernière partie de ce manuscrit est donc une présentation des résultats récemment obtenus au laboratoire avec ces interacteurs et avec la bibliothèque d'*αRep* de première génération décrite dans le Chapitre I ainsi que la banque de deuxième génération 2.1 décrite dans le Chapitre II.C.

E. Travaux supplémentaires

Introduction

Des travaux supplémentaires ont été réalisés avec la banque d'*αRep* de deuxième génération 2.1.

En effet, les membres de l'équipe ont réalisés d'autres sélections pour identifier des interacteurs spécifiques de cibles qui sont soit des cibles modèles permettant de valider la stratégie soit des cibles avec des applications potentielles. Les deux banques d'*αRep* : de première (1.0) et de deuxième génération 2.1 ont été utilisées.

1. Les sélections effectuées avec la banque d'αRep de première génération 1.0

Bien qu'elle n'ait pas été conçue expressément pour cela et qu'elle ne soit pas réellement optimisée la banque d'*αRep* de première génération1.0 comportait une diversité correcte. Des sélections ont été réalisées au laboratoire avec cette banque et des variants qui interagissent avec différentes cibles ont été identifiés et les interactions caractérisées : (Résultats non publiés).

On s'est proposé de chercher des interacteurs qui pourraient reconnaitre la face constante et commune à toutes les *αRep*, ce qui aurait constitué un outil intéressant pour la filtration des clones non codants de la banque de seconde génération. L'idée qui a été suivie, consiste à utiliser l'α-Rep-n4-a comme cible pour des sélections avec la banque de première génération exposée sur phages (*phage display*). L'α-Rep-n4-a a été utilisée comme cible, car c'est l'*αRep* dont la structure résolue montre qu'elle forme un dimère par sa surface rendue variable. L'hypothèse était donc d'immobiliser le dimère et par la suite le mettre en contact avec les phages exposant les *αRep*. L'idée était d'identifier des variants qui reconnaissent la face exposée dans le dimère qui correspond à la face commune et constante de toutes les protéines («le dos »). Ces sélections ont été réalisées comme décrit au préalable (Chapitre II-D) et elles ont permis d'identifier plusieurs interacteurs spécifiques. Parmi ces interacteurs caractérisés, nous avons choisi le clone α-Rep-n1-a (α2) pour résoudre la structure du complexe (Fig. 1).

	Cible /interacteurs	α-Rep-n4-a	
		K_D	n
Bq *αRep* I^aire^ génération	α-Rep-n3-a (α1)	**1.11± 0.58µM**	**0.732±0.043**
	α-Rep-n1-a (α2)	**15.8± 10Nm**	**0.861±0.007**

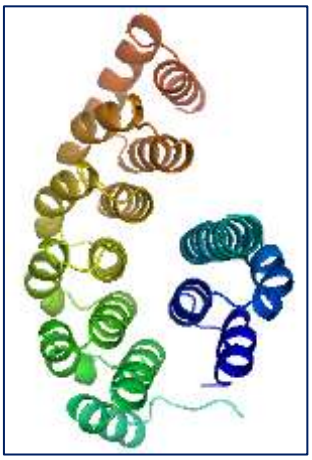


Fig. 1 : La structure résolue pour le complexe [α-Rep-n4-a/ α-Rep-n1-a (α2)] en collaboration avec l'équipe FAAM.

Cette structure montre clairement que les *αRep* : α-Rep-n4-a et α-Rep-n1-a (α2) interagissent par leurs surfaces rendues variables, et non pas non comme ce qui était attendu par la face variable de α-Rep-n1-a (α2) avec le « dos » du dimère de α-Rep-n4-a. Il apparait que le dimère de l'α-Rep-n4-a se dissocie pour interagir avec l'autre *αRep*. Par ailleurs, dans le cristal, α2 voit son *Ccap* clivé, probablement par des traces de protéases présentes dans l'échantillon utilisé pour la cristallogenèse. Le fragment de *Ccap* non clivé ne reste pas à la place attendue pour un *Ccap* mais devient un prolongement de l'hélice 2 du module précèdent. Cela conduit à une extension d'hélice et à l'apparition d'une surface disponible pour contracter des interactions inter-hélices avec une molécule voisine. Ce processus connu sous le nom d'échange de domaine (*Domain swapping*) est possible avec de nombreux autres types de protéines modulaires et revient à échanger des interactions originellement intramoléculaires en interaction intermoléculaires. Le déclencheur est ici probablement le clivage artefactuel d'une partie du *Ccap*.

On peut également raisonner, *a posteriori*, qu'un interacteur *αRep*, interagissant avec le « dos » d'autres *αRep*, interagirait potentiellement aussi avec lui-même. Un tel objet serait de ce fait une protéine qui s'oligomeriserait, ce qui pourrait peut-être compliquer le processus d'exposition sur phages et la possibilité de le sélectionner. Dans ce contexte, nous avons choisi, comme cela a été exposé, de ne pas prolonger cette stratégie et de filtrer plutôt la banque par le biais du *Flag-tag*.

Des sélections ont été aussi réalisées sur un variant d'une banque de Néocarzinostatine (Chapitre II-A) : la NCS3.24. Cette sélection a permis d'identifier une *αRep* interagissant avec la cible. Bien que la sélection n'ait pas été orientée dans ce sens cet interacteur n'interagit pas

avec la forme sauvage de la NCS qui ne diffère de la NCS3.24 que par moins de 10 % de sa séquence. La structure du complexe a été résolue par l'équipe de H. Van Tilbeurgh et est représentée et sur la Fig. 2. Ce complexe montre que la zone de contact entre les partenaires implique la face variable de l'αRep et la région comportant les chaines latérales de la NCS 324 qui sont mutées relativement à la protéine sauvage. Cette région forme une crevasse riche en chaines latérales aromatiques et sa composition particulière peut en avoir fait un site favorable à l'établissement d'interaction.

	Cible /interacteurs	**NCS324**	
		K_D	n
Bq *αRep* 1ère génération	α-Rep-n2-a (cln 16)	**0.94± 0.22 µM**	**1.43±0.029**

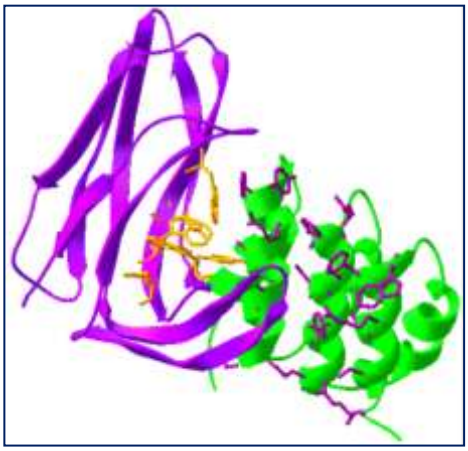


Fig. 2 : La structures du complexe [α-Rep-n2-a (cln 16)/NCS 324] résolue en collaboration avec l'équipe FAAM.

Ces résultats sont très encourageants et suggère qu'il est possible d'identifier des interacteurs à partir de l'ossature *αRep*. La banque de première génération, bien qu'elle ne soit pas optimale, permet déjà de sélectionner des interacteurs ayants des constantes de dissociation comprises entre 10^{-8} et 10^{-6} M. Les interacteurs obtenus avec la bibliothèque de première génération ont tous 2 ou 3 modules répétés, ce qui résulte probablement du fait que les séquences plus longues sont rarement correctes dans cette bibliothèque.

Les sélections avec la banque de seconde génération va probablement conduire à des interacteurs possédant une diversité plus élevée et un plus grand nombre de modules. Les travaux réalisés en collaboration ont permis assez rapidement l'obtention de structures, et les

résultats de cristallogenèse suggèrent que ces protéines seules ou complexées semblent cristalliser assez facilement. Enfin les cibles reconnues peuvent avoir des types de structures secondaires ou tertiaires très différentes. Les deux structures obtenues montrent en effet que le type d'objet reconnu n'est pas limité aux protéines ayant une structure hélicoïdale. Par ailleurs, la reconnaissance est réellement spécifique et peut discriminer des variants proches l'un de l'autre.

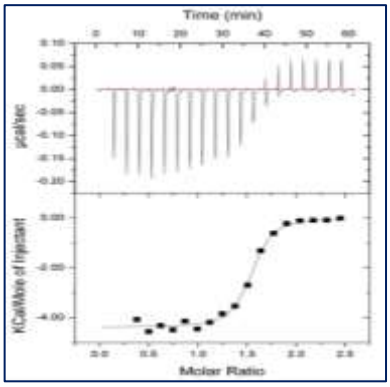
2. Les sélections effectuées avec la banque d'αRep de deuxième génération 2.1

Les résultats préalables nous ont encouragés à générer d'autres outils. Des sélections à partir de la banque de deuxième génération 2.1 ont été réalisées d'une part pour la cible Fibronectin-Binding Protein (FNE) dans le but de favoriser sa cristallogenèse et d'autre part pour de petites GTPases *Arf* pour l'utilisation comme traceurs intracellulaires.

2.1. Résultats obtenus par sélections avec la cible FNE

La FNE, étudiée par l'équipe FAAM, est une protéine de *Streptococcus equi* qui interagit avec la fibronectine. Ces protéines ne sont pas très bien caractérisées mais elles semblent être impliquées dans l'adhésion et la virulence des bactéries via leur interaction avec la fibronectine au niveau de la matrice extracellulaire de la cellule hôte. Elles ont un effet déterminant dans l'augmentation du potentiel pathogène des bactéries. Cette sous espèce de Streptococcus est connue pour être un pathogène des chevaux chez qui elle cause une maladie contagieuse des voies respiratoires supérieures appelée la gourme.

L'objectif de nos collaborateurs était de comprendre ces interactions et pour cela de résoudre la structure de la FNE par cristallographie. Aucune tentative pour cristalliser la FNE toute seule n'aboutit. Nous nous sommes ainsi proposés de générer des *αRep* spécifiques de cette cible qui pourraient aider cette protéine à cristalliser et permettre de résoudre la structure par la suite. Des sélections ont été réalisées et plusieurs interacteurs ont été identifiés et les interactions caractérisées. Le variant le plus affin à la FNE a une constante de dissociation mesurée par ITC $K_D = 0{,}178.10^{-6} \pm 0{,}036.10^{-6}$ M (Fig. 3).



N =1,49 ± 0,010
K_D = 0,178.10^{-6} ± 0,36 * 10^{-7} M

Fig. 3 : Thermogramme de l'interaction FNE-*αRep3*.

Le complexe a été purifié, concentré puis cristallisé. Et la structure du complexe a été résolue (Fig. 4) dans le cadre de la thèse de Mourira Tiouajni « ***Vers une meilleur compréhension de la matrice extracellulaire : Etude structurale et fonctionnelle des complexes entre la fibronectine et différents partenaires »*** élaborée sous la sous la direction H. Van Tilbeurgh.

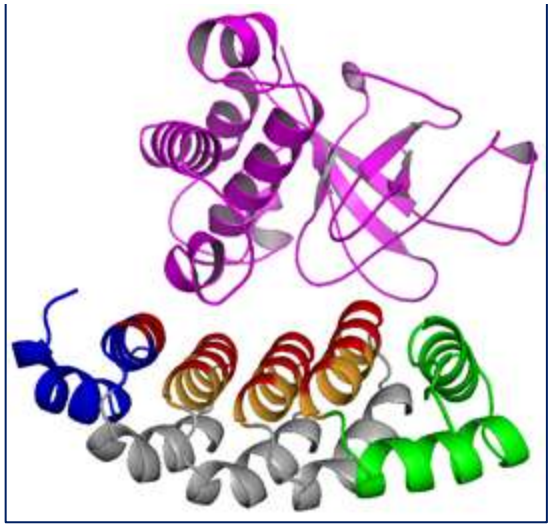


Fig. 4 : Structure résolue du complexe FNE-*αRep3* : La FNE est représentée en violet et l'*αRep3* est une protéine à 3 motifs répétés dont le *Ncap* est représenté en bleu, le *Ccap* en vert et la surface d'interaction est représentée en rouge.

Cette structure montre clairement que l'*αRep3* interagit avec sa surface rendue variable. La Fig. 5 représentée ci-dessous illustre les liaisons et les résidus impliqués dans l'interaction.

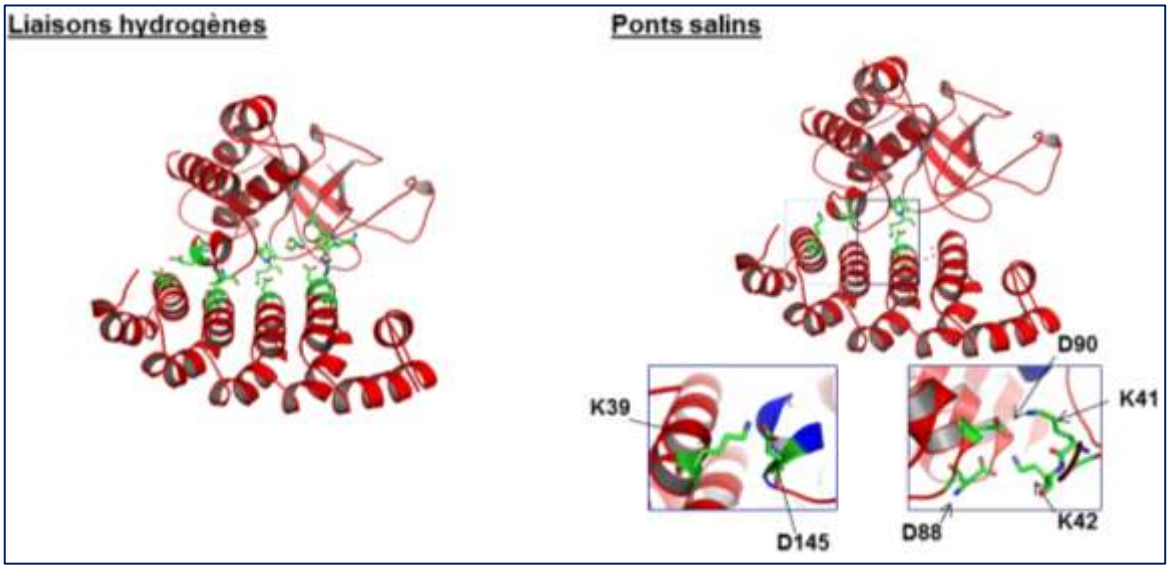


Fig. 5 : Les liaisons et les résidus impliqués dans l'interaction FNE-αRep3.

Les résidus impliqués dans l'interaction sont ceux des positions hypervariables y compris les positions portées par le *Ncap*, ce qui confirme qu'il peut être utile d'exploiter ces positions. Les interactions mises en œuvre pour l'interaction sont essentiellement des liaisons hydrogènes et des ponts salins qui sont représentés sur la figure ci-dessous (Fig. 6). 7 résidus parmi des 24 de l'*αRep* interagissent avec la FNE alors que la 8^{éme} liaison est assuré par le résidu D de la position 17 constante du deuxième motif.

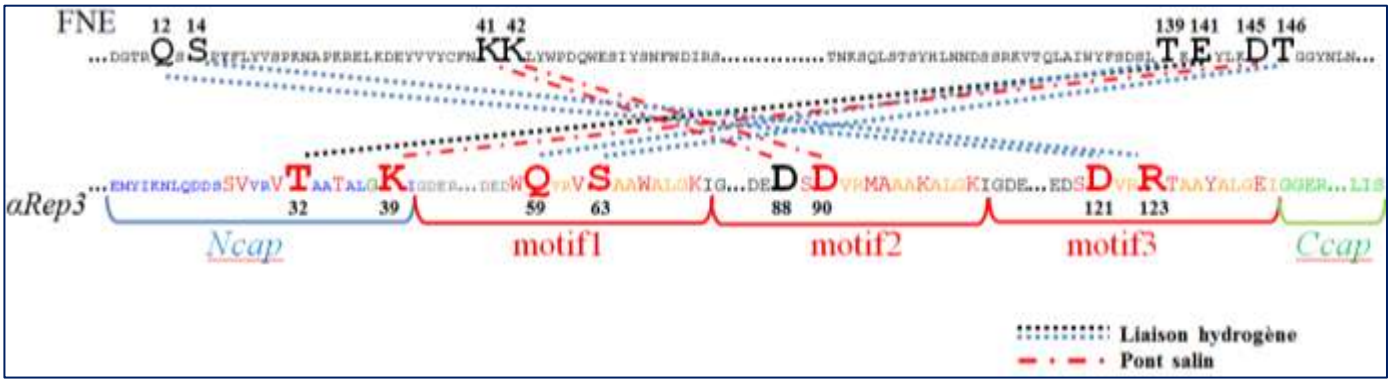


Fig. 6 : Les liaisons établies lors de l'interaction αRep3 - FNE

Ce travail démontre que la génération d'interacteurs spécifiques peut permettre de résoudre des structures cristallographiques qui résistaient jusque-là aux essais de cristallisation

2.2. Les résultats obtenus par sélections pour les petites GTPases *Arf*

Le laboratoire s'est proposé de rechercher au sein de notre banque de protéines artificielles des variants qui soient capables d'interagir spécifiquement avec des protéines étudiées pour leur rôle dans la signalisation les protéines *Arf* et de discriminer deux isoformes de petites GTPases (*Arf1* et *Arf6*). L'objectif était également de voir s'il était possible de les discriminer en fonction de leur état fonctionnel (actif/inactif). Ce projet a été mené en collaboration avec l'équipe de Jacqueline Cherfils du Laboratoire d'Enzymologie et Biochimie Structurales (LEBS). L'intérêt de cette collaboration est de créer les tous premiers outils de caractérisation de l'état d'activation et la localisation subcellulaire de ces interrupteurs moléculaires à l'échelle cellulaire, en particulier en absence à ce jour d'anticorps ou de réactifs ayant ces capacités de reconnaissance.

Les deux isoformes des petites GTPases pris comme cibles sont *Arf1* et *Arf6*. La première est la plus abondante des petites protéines G et elle est impliquée principalement dans le transport vésiculaire de l'appareil de Golgi. L'*Arf6* est plutôt localisée au niveau de la membrane plasmique et intervient dans la coordination entre le trafic membranaire et la

dynamique du cytosquelette à la périphérie de la cellule. L'alternance structurale entre la forme active (GTP) et la forme inactive (BDP) a été bien étudiée par cristallographie et par RMN notamment par l'équipe de Jacqueline Cherfils (Biou et *al.*, 2010). Les deux formes GTP et GDP d'une même isoforme sont différentes entre elles et le remodelage structural va permettre aux GTPases *Arf* de coupler leur recrutement à la membrane à leur activation par échange GDP/GTP.

Des sélections indépendantes ont été réalisées, en prenant comme cibles les deux formes GDP et GTP d'*Arf1* et les 2 formes GDP et GTP d'*Arf6*.

Plusieurs interacteurs ont été identifiés par phages Elisa. Ces variants ont été produits puis purifiés en grandes quantités et l'interaction a été caractérisée par Elisa protéine – protéine. Ces tests ont permis d'étudier la spécificité de la reconnaissance entre isoformes et même entre formes active et inactive (Fig. 8).

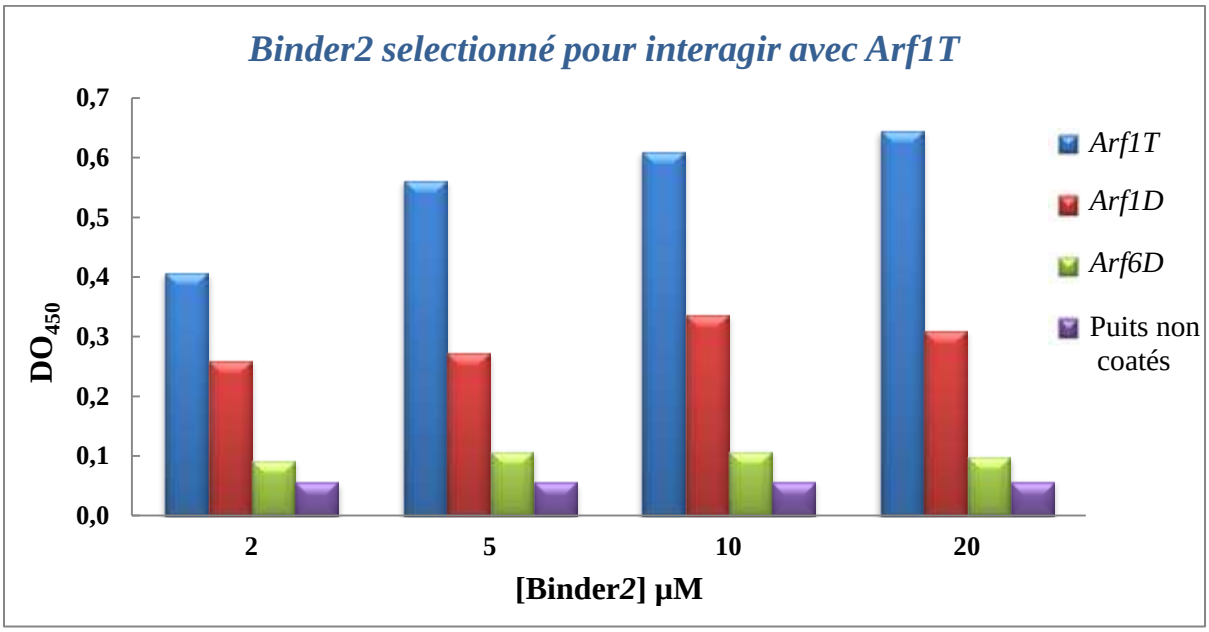


Fig. 8 : Test Elisa protéine – protéine du *Binder2* avec les *Arf1T* (en bleu), *Arf1D* (en rouge) et *Arf6D* (en vert).

Ce test Elisa montre une nette spécificité de la reconnaissance entre les deux isoformes (*Arf1* et *Arf6*). La spécificité est moins évidente pour les formes active et inactive (histogrammes bleus et rouges) toutefois il est clair que l'affinité pour *Arf1T* est plus importante que celle qu'il a pour *Arf1D*. Ce test Elisa bien qu'il fournisse des informations importantes, reste qualitatif et les conclusions doivent être confirmées par mesure de la constante de dissociation. Ces travaux sont en phase de finalisation par les membres de notre équipe.

Ces travaux montrent les caractéristiques et les capacités d'affinité et de spécificité dont nos protéines artificielles sont dotées et nous encouragent à exploiter cette ressource, pour créer des protéines nouvelles utiles dans le cadre de projets collaboratifs orientés d'une part sur des approches de biologie structurale des protéines membranaires, et d'autre part sur la perturbation in vivo d'interfaces protéine/protéine choisies, ou plus généralement sur des stratégies d'interférences par protéines.

Conclusion et perspectives

Conclusion

Le travail de thèse nous a permis de construire une banque de protéines artificielles basée sur la répétition d'un motif particulier en double hélice α. Trois banques ont été construites au laboratoire:

- Une première banque : ***banque αRep première génération1.0*** : Cette banque nous a permis de valider la conception de l'architecture choisie ainsi que la possibilité de construire efficacement des banques : ***Ncap*-(motif)$_n$-*Ccap***. Ma contribution à ce travail entamé avant mon arrivée au laboratoire a porté sur la caractérisation de cette banque et de ses protéines.
- Une deuxième banque : ***banque αRep deuxième génération 2.0 :*** Cette banque nous a permis de tester les améliorations apportées au vecteur accepteur (inversion des promoteurs et ajout d'un outil de filtration), à la construction moléculaire de la banque (utilisation d'oligonucléotides de cercles double brin) et finalement à la distribution des acides aminés aux positions hypervariables.
- Une troisième banque : ***banque αRep deuxième génération 2.1*** obtenue par un processus de construction impliquant plusieurs étapes. Une étape d'assemblage, une étape de filtration et une étape de recombinaison de modules ou S*huffling*. Cette banque est la plus diverse et comporte uniquement des modules codants.

Des sélections ont été réalisées contre une protéine modèle (la TEV protéase) et 2 autres protéines cibles dans le but d'obtenir des outils d' « aide à la cristallographie ». Des interacteurs ont été identifiés et les méthodologies de sélection ainsi que de tri des clones positifs ont été développés. Ce qui a permis de montrer clairement que pour chacune de ces cibles des interacteurs spécifiques peuvent effectivement être produits. Les affinités des protéines pour leur cible ne sont pas encore caractérisées.

D'autres sélections ont été réalisées en parallèle par les membres de l'équipe avec différentes autres cibles. Ceci nous a permis de vérifier qu'il est possible de sélectionner des interacteurs spécifiques pour des cibles ayant des caractéristiques structurales diverses. La gamme d'affinité des interacteurs obtenus est correspond à des K_D compris entre 10^{-8} et 10^{-6}, ou de quelques nM pour les plus affins actuellement détectés à partir de la banque 2.1. Il

semble donc établi que cette technologie atteint son but premier qui était de permettre de rechercher efficacement des interacteurs spécifiques contre des protéines cibles variées.

Perspectives

Notre laboratoire dispose maintenant d'une ressource générique pour produire de nouvelles protéines permettant d'aborder des projets divers. En effet, ces protéines peuvent être utilisées dans bon nombre d'applications qui mettent en œuvre actuellement les anticorps monoclonaux, avec l'avantage d'être produites en grande quantité, plus simplement, en n'ayant recours ni à l'immunisation d'animaux ni aux cultures de cellules animales.

Il parait important d'améliorer le procédé de sélection par *phages display* pour permettre de sélectionner plus rapidement des interacteurs recherchés. Une fois la bibliothèque construite, l'une des étapes limitantes en termes de temps expérimental devient la production des cibles purifiées nécessaire aux sélections. Des améliorations permettant de conduire les sélections à partir de protéines non purifiées mais biotinylées *in vivo* ont été testées au laboratoire et permettent probablement de simplifier cette étape. Le système d'interaction Biotine-Streptavidine permet également de garder des cibles dans leur conformation native et d'éviter toute dégradation de la qualité de la cible due à l'adsorption de celle-ci à la surface de la plaque de sélection. Les étapes de détection et caractérisation des clones positifs en sortie de sélection seront également longues à mettre en œuvre si chaque séquence positive doit être sous clonée et produite en grande quantité. Les tests réalisés dans cette thèse avec des protéines produites directement à partir du vecteur dans des souches d'expression et utilisant des extraits solubles non purifiés permettent de simplifier ces cribles secondaires et de réserver les tests les plus longs aux clones les plus performants.

Selon les applications envisagées, nous pouvons aussi chercher à orienter le processus de sélection de façon à ne sélectionner que les variants ayant de bonnes affinités pour leurs cibles (nM). Ceci peut être réalisé de deux façon différentes soit en allongeant les temps de lavages, ce qui va permettre de ne conserver puis d'éluer que les variants les plus affins, soit en ayant recours à un processus de *maturation d'affinité*. Cette dernière stratégie consiste à sélectionner dans un premier temps des variants plus ou moins affins à une cible donnée puis dans un second temps à les améliorer. La nature modulaire de l'architecture *αRep*

permet de rechercher de nouvelles combinaisons de modules il parait vraisemblable qu'un interacteur de première génération puisse être amélioré par l'apport de modules additionnels. Pour réaliser cela expérimentalement, les motifs constituants un interacteur initial peuvent être extraits puis combinés avec une proportion minoritaire de modules aléatoires dans le but de recréer de la diversité et d'explorer plus complètement autour d'une solution initiale. De nouveaux tours de sélection permettront de rechercher parmi les nouveaux variants ceux qui ont des affinités plus importantes que celles des interacteurs initiaux.

Une fois la stratégie de sélection améliorée, nous pouvons envisager d'utiliser nos protéines pour des applications variées. Les applications en biologie structurale de ce type d'objet sont très vraisemblables. Les interacteurs spécifiques de cibles variées peuvent être sélectionnés et ensuite produits en quantité suffisante pour aborder les études structurales. Les protéines couvrant une partie importante de la surface de la cible peuvent moduler très notablement d'une part sa stabilité et sa dynamique et d'autre part le processus de cristallogenèse lui-même en changeant de façon importante les possibilités de contacts intermoléculaires. Les résultats déjà obtenus au laboratoire suggèrent que cette stratégie doit être explorée en particulier pour des problèmes difficiles de biologie structurale comme la détermination de la structure de protéines membranaires. Les résultats obtenus avec les *Darpins* ou d'autres ossatures montrent que cette approche peut être réellement efficace pour cristalliser des protéines membranaires. A titre d'exemple, des « *nanobodies* » (fragments d'anticorps de camélidés) ont été générés pour interagir avec le récepteur adrénergique β_2 humain (β_2 AR). Un de ces « *nanobodies* » a montré des comportements similaires aux protéines G et même un comportement agoniste ce qui a permis la résolution de la structure du récepteur à l'état actif (Rasmussen et *al.*, 2011).

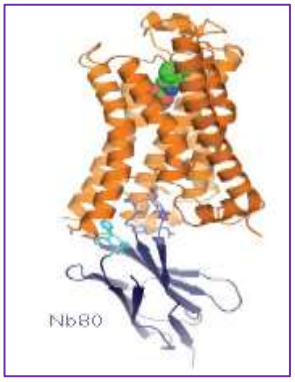


Fig. 1 : Structure résolue du complexe β_2AR-Nb80 : β_2AR est représenté en orange et le « *nanobody* » Nb80 en bleu. Les CDRs du « *nanobody* » impliquées dans l'interaction sont présentés en bleu clair.

Ces protéines artificielles peuvent être dotées, en plus des capacités de reconnaissance et d'interaction, d'un pouvoir effecteur sur l'activité des protéines qu'elles fixent. Plückthun et ses collaborateurs sont arrivés à identifier, à partir de leurs bibliothèques d'ankyrine (*DARPins*), des variants qui peuvent reconnaitre trois épitopes très similaires d'un transporteur ABC : LmrCD de *Lactococcus lactis*. Et par un test fonctionnel, dans *Lactococcus lactis*, ils ont pu mettre en évidence trois *DARPins* qui permettent d'activer le transport de médicaments par une augmentation de l'activité du LmrCD suite à l'interaction avec l'ankyrine (Seeger et *al.*, 2012).

Plusieurs protéines artificielles à potentiel thérapeutique ont été développées et sont même en phase 2 du développement clinique (Fig2)[5]. Si la sélection d'*αRep* spécifique de cible thérapeutique potentielle est certainement concevable, il n'est pas établi que les *αRep* trouveront là un champ d'application important. Les applications thérapeutiques supposent en effet que le caractère immunogène de ces protéines chez l'homme ne soit pas un obstacle majeur. Les séquences d'*αRep* n'ont pas été conçues avec le souci de minimiser leur immunogénicité potentielle et si des protéines de type *HEAT repeats* existent chez l'homme, le sous-groupe de séquence utilisé ici est plus représenté chez les microorganismes surtout thermophiles. Aucune indication sur leur éventuel caractère immunogène n'est actuellement disponible.

Enfin, et une application importante pour ces protéines est l'utilisation de variants spécifiques à une protéine cytoplasmique donnée pour évaluer les conséquences de sa séquestration dans une cellule vivante. Ces travaux peuvent plus difficilement être réalisés avec des anticorps puisque ceux-ci sont souvent difficiles à exprimer efficacement dans des conditions réductrices. Ces approches de *phenotypic knock-out* ou ***protéine-interférence*** sont faiblement explorées fautes de technologie performante pour produire les protéines nécessaires. De ce point de vue les travaux décrits ici pourraient utilement complémenter utilement les travaux visant à concevoir, produire et mettre en œuvre des « *intrabodies* » (Butler, McLear, and Messer, 2012).

[5] http://molecularpartners.com/public/index.php?id=105&lang=en

Références bibliographiques

Biou, Valérie, Kaheina Aizel, Pierre Roblin, Aurélien Thureau, Eric Jacquet, Sebastian Hansson, Bernard Guibert, et al. 2010. "SAXS and X-ray Crystallography Suggest an Unfolding Model for the GDP/GTP Conformational Switch of the Small GTPase Arf6." *Journal of Molecular Biology* 402 (4) (October 1): 696–707. doi:10.1016/j.jmb.2010.08.002.

Butler, David C., Julie A. McLear, and Anne Messer. 2012. "Engineered Antibody Therapies to Counteract Mutant Huntingtin and Related Toxic Intracellular Proteins." *Progress in Neurobiology* 97 (2) (May): 190–204. doi:10.1016/j.pneurobio.2011.11.004.

Luke, Brian, Claus M Azzalin, Nele Hug, Anna Deplazes, Matthias Peter, and Joachim Lingner. 2007. "Saccharomyces Cerevisiae Ebs1p Is a Putative Ortholog of Human Smg7 and Promotes Nonsense-mediated mRNA Decay." *Nucleic Acids Research* 35 (22): 7688–7697. doi:10.1093/nar/gkm912.

Rasmussen, Søren G. F., Hee-Jung Choi, Juan Jose Fung, Els Pardon, Paola Casarosa, Pil Seok Chae, Brian T. DeVree, et al. 2011. "Structure of a Nanobody-stabilized Active State of the [bgr]2 Adrenoceptor." *Nature* 469 (7329) (January 13): 175–180. doi:10.1038/nature09648.

Seeger, Markus A., Anshumali Mittal, Saroj Velamakanni, Michael Hohl, Stefan Schauer, Ihsene Salaa, Markus G. Grütter, and Hendrik W. van Veen. 2012. "Tuning the Drug Efflux Activity of an ABC Transporter in Vivo by in Vitro Selected DARPin Binders." Ed. Dan Zilberstein. *PLoS ONE* 7 (6) (June 4): e37845. doi:10.1371/journal.pone.0037845.

Soltes, Glenn, Heather Barker, Kristine Marmai, Elaine Pun, Amy Yuen, and Erik J Wiersma. 2003. "A New Helper Phage and Phagemid Vector System Improves Viral Display of Antibody Fab Fragments and Avoids Propagation of Insert-less Virions." *Journal of Immunological Methods* 274 (1-2) (March 1): 233–244.

Urvoas, Agathe, Asma Guellouz, Marie Valerio-Lepiniec, Marc Graille, Dominique Durand, Danielle C. Desravines, Herman van Tilbeurgh, Michel Desmadril, and Philippe Minard. 2010a. "Design, Production and Molecular Structure of a New Family of Artificial Alpha-helicoidal Repeat Proteins (αRep) Based on Thermostable HEAT-like Repeats." *Journal of Molecular Biology* 404 (2) (November 26): 307–327. doi:10.1016/j.jmb.2010.09.048.

———. 2010b. "Design, Production and Molecular Structure of a New Family of Artificial Alpha-helicoidal Repeat Proteins ([alpha]Rep) Based on Thermostable HEAT-like Repeats." *Journal of Molecular Biology* 404 (2) (November 26): 307–327. doi:16/j.jmb.2010.09.048.

Virnekäs, B, L Ge, A Plückthun, K C Schneider, G Wellnhofer, and S E Moroney. 1994. "Trinucleotide Phosphoramidites: Ideal Reagents for the Synthesis of Mixed Oligonucleotides for Random Mutagenesis." *Nucleic Acids Research* 22 (25) (December 25): 5600–5607.

Solid-phase Synthesis and Stabilization of mRNA–protein Fusions" 37 (16) (September): e108–e108. doi:10.1093/nar/gkp514.

Annexe

Cette annexe résume des procédures expérimentales suivies pour la réalisation de ce travail.

Sous-clonage de la séquence d'une protéine d'un vecteur à un autre vecteur

Le sous-clonage par le biais de deux enzymes de restriction X et Y.

- *Digestion des deux plasmides : avec la séquence de la protéine et le nouveau vecteur avec les 2 enzymes de restriction : incubation à la température d'activation des 2 enzymes pendant 1 h.*
- *Dépôt sur gel d'agarose 2 % le produit de la double digestion du vecteur contenant la protéine et sur gel d'agarose 1 % le produit de la double digestion du vecteur accepteur.*
- *Purification des bandes correspondant à la région codant la protéine et celle correspondant au vecteur :*
  - *Découpe du gel*
  - *Purification de l'ADN à partir du gel par le kit ExtractII de chez Macherey-Nagel.*
- *Ligation des 2 fragments o.n. à 16°C.*
- *Purification du produit de la ligation puis transformation de bactéries XL1Blue MRF' électrocompétentes.*
- *Repiquage de clones ont été pris au hasard pour vérifier l'intégration de l'insert (séquençage).*

Amplification de cercles d'ADN avec des amorces synthétisés: « *Amplification home made* »

- *Mélanger les cercles d'ADN avec le tampon de la Φ29, les dNTP et les primers synthétiques.*
- *Chauffage 3 min à 95°C puis refroidissement jusqu'à 4°C.*
- *Ajout de 5 unités de Φ29 polymérase et la pyrophosphatase ainsi que la BSA.*
- *Incubation 4 h à 3°C.*
- *Analyse du produit d'amplification.*

Amplification des cercles du motif consensus avec le kit TempliPhi™ de GE:

- *Mélanger les cercles (1 µL) avec le « Sample Buffer » du kit (10µL).*
- *Chauffage à 95°C pendant 3 min puis refroidissement à 4°C.*
- *Ajout de l'«Enzyme mix» (0.4 µL)*
- *Ajout de 10 µL du «Reaction buffer*
- *Incubation 4h à 30°C.*
- *Analyse du produit d'amplification :*
  - *Ajout de 240 µL d'eau pour arrêter la réaction.*

- *Digestion par une enzyme de restriction : une digestion partielle (0.2 µL d'enzyme incubé 10 min) et une digestion totale (1 µL d'enzyme incubé 1 h)*

Purification des produits de digestion et dépôt sur gel d'agarose.

Test d'expression en milieu liquide

Le test d'expression en milieu liquide nécessite deux étapes :

- Expression des protéines
- Analyse par *Western Blot*

✓ *Expression des protéines*

- *Transformation de bactéries BL21DE3 chimio- compétentes avec les constructions plasmiques à tester (voir Expression et Purification comparative de la NCS).*
- *Incubation d'une colonie dans 10 mL 2YT+Amp+Glu o.n. à 37°C et 220rpm : préculture.*
- *Inoculation de cultures de 20 mL 2YT+Amp à DOi=0.1 et incubation à 37°C et 220 rpm.*
- *A DO_{600}= 0.6, induction de l'expression par l'ajout d'IPTG : 4h à 37°C et 220 rpm.*
- *Récupération des bactéries par centrifugation (10', 4000 g).*
- *Lyse des bactéries par le tampon B-PERII (Roche).*
- *Prélèvement d'échantillons du lysat total et de la fraction soluble puis dénaturation.*

✓ *Analyse par Western Blot*

- *Dépôt des échantillons dénaturés sur gel acrylamide 15% et séparation des protéines par électrophorèse (150 V et 40mA/gel pendant 50 min).*
- *Transfert des protéines sur une membrane de nitrocellulose (150V et 300 mA pendant 45 min)*
- *Blocage de la membrane au TBST BSA 3% sous agitation : 2 h à température ambiante.*
- *Révélation de la membrane :*
  - *Incubation dans 10 mL d'une solution de TBST avec un anticorps anti-His-tag (dilution 1/10000).*
  - *3 lavages au TBST*
  - *Incubation dans 10 mL une solution de TBST avec l'anticorps secondaire : anticorps anti-souris fluorescent (dilution 1/50000).*
  - *Membrane scannée dans le scanner Odyssée à une longueur d'onde d'excitation de 700nm.*

Test d'expression soluble en milieu solide : *Cofiblot*

Ce test permet de tester l'expression soluble de 72 clones à la fois.

- *Transformation des souches d'expression d'E. coli (Rosetta Blue ou BL21DE3) par une collection de phagemides et étaler sur boîtes de 2YTagar + Amp+Glu.*
- *Remplissage d'une plaque de culture 96 puits avec 150 µL 2YT + Amp+Glu.*

- *Inoculer 1 colonie transformée par puits de la plaque précédente : incubation o.n. à 37°C à 500 rpm.*
- *Repiquage de la plaque 96 puits, par un peigne, sur filtre Durapore déposé sur milieu solide 2YTagar+Amp o.n. à 37°C.*
- *Incubation du même filtre sur milieu solide 2YTagar+Amp+IPTG 4h à 37°C.*
- *Lyse des colonies et transfert des protéines solubles sur membrane de nitrocellulose:*
    - *Dépôt du filtre Durapore, colonies vers le haut, sur un « sandwich » composé d'un papier Wattman et d'une membrane de nitrocellulose.*
    - *Le sandwich doit être imbibé de 5 mL de tampon de lyse B-PERII : 30 min à température ambiante.*
- *Révélation de la membrane de nitrocellulose (Test d'expression en milieu liquide)*

Expression et Purification comparative de la NCS

- *Transformation de bactéries BL21DE3 chimio-compétentes (250 µL) :*
    - *Incuber les bactéries 30 min sur glace en présence du plasmide.*
    - *Incuber 45 secondes à 42°C.*
    - *Incuber 2 min sur glace.*
    - *Ajouter 200 µL de milieu 2YT et incuber 1 h à 37°C et 80 rpm.*
    - *100 µL des bactéries transformées sont étalés sur une boîte de 2YTagar+Amp+Glu et incubés o.n. à 37°C.*
- *Une colonie de la boîte est utilisée pour préparer une préculture 50 mL 2YT+Amp+Glu incubée o.n. à 37°C.*
- *La DO de la préculture est mesurée dans le but d'ensemencer 500 mL 2YT+Amp à DOi = 0.1.*
- *La culture est incubée 72h à 30°C : Expression par auto-induction puis sécrétion dans le milieu de culture.*
- *Récupération du surnagent de culture par centrifugation (20', 4000g).*
- *Purification de la protéine à partir du 2 surnageant :*
    - *Précipitation au sulfate d'ammonium.*
    - *Solubilisation du culot protéique dans l'eau puis dialyse*
    - *Purification par une étape de chromatographie d'affinité sur résine de Nickel.*
- *Analyse des différentes fractions de la purification par SDS page sur gel d'acrylamide 15%.*

Production de phages à partir d'un clone

- *Transformation de bactéries XL1Blue MRF' avec la construction plasmique d'exposition sur phages.*
- *Incubation d'une colonie o.n. à 37°C et 220 rpm dans 10 mL 2YT+Amp+Tet+Glu.*
- *Préparation d'une culture de 20 mL 2YT+Amp+Tet à DOi=0.1.*
- *A DO=0.6, infection des bactéries par le phage helper (avec une multiplicité de 10) et incubation à 37°C pendant 30min sans agitation puis 30 min avec une agitation de 80 rpm. Le volume de phage helper est calculé selon la formule :*

*$(5*10^8*DO*V_{culture}$*facteur de multiplicité) / (Titre du phage hepler)*

- *Récupération des bactéries infectées centrifugation (10 min, 4000 g).*
- *Suspension des bactéries dans 20 mL 2YT+Amp+Kan et incubation o.n à 30°C à 220 rpm en présence ou en absence d'IPTG.*
- *Récupération des phages dans le surnagent de culture par centrifugation (30 min, 8000 g).*
- *Concentration et purification des phages par précipitation au PEG :*
  - *Mélange surnageant-PEG NaCl 1:1 (V/V) : 20 % PEG 8000, 25M NaCl pendant 20min à température ambiante.*
  - *Récupération des phages précipités par centrifugation (30min, 12000 g).*
  - *Suspension dans TBS et précipitation une $2^{éme}$ fois.*
- *Suspension des phages dans TBS.*
- *Dosage des phages*
  - *Infection de bactéries XL1Blue MRF' :*
    - *Préparation des dilutions de phages : 50 µL de phages dans 450 µL de 2YT.*
    - *Préparation d'une culture de XL1Blue MRF' dans 2YT+Tet à DO = 0.6*
    - *Mélange et incubation 15 min à 37°C et 80 rpm : 50µL de la dilution de phages avec 4500µL de XL1Blue MRF' à DO=0.6.*
    - *Etaler 50µL de bactéries infectées étalés sur du milieu solide 2YT agar+Amp+Glu.*
    - *Comptage des colonies sur boîtes, ont été comptées et détermination de la concentration de phages dans la solution.*

Les colonies qui poussent sur les boîtes sont comptées et elles reflètent un nombre de phages infectibles : ***nombre de colonies *10 *20*10^{dil} (cfu/mL).***

- *Mesure de la DO à 269 nm et estimation de la concentration en particule de phage selon la formule suivante :*

$(DO_{269}*6*10^{16})$ / (Nombre de pb du phagemide)

Test phage Elisa en pool

- *Les phages représentant la banque et les différents tours de sélection sont prélevés des suspensions de phages dialysés utilisées aux différents tours de sélection.*
- *Immobilisation des cibles (voir Tour de sélection)*
- *3 Lavages TBST.*
- *Blocage avec 250 µL de TBST BSA 3% : 3 h à 15°C.*
- *3 lavages au TBST.*
- *Incubation avec 100 µL de chaque échantillon de phages : 2 h à 25°C.*
- *4 lavages au TBST.*
- *Révélation de la présence des phages reconnaissant la cible immobilisée :*
  - *Incubation 1 h à 20°C avec 100 µL d'une solution de TBST avec l'anticorps anti-M13 couplé à la peroxydase à la dilution 1/5000.*
  - *3 lavages au TBST*

- *Ajouté de 100 µL de substrat soluble.*
- *Apparition d'une coloration bleue qui s'intensifie avec le temps d'incubation.*
- *Incubation 5 min d'incubation sous agitation.*
- *Arrêter la réaction par ajout de 100 µL d'acide HCl 1 N.*

Test phage Elisa Clonal

A l'issus des différents tours de sélection, nous obtenons des boîtes avec des colonies représentatives du tour en question. A partir de ces colonies, on prépare une plaque 96 puits de culture où chaque ligne va correspondre à un tour particulier (***« plaque référence »***).

✓ Production des phages clonaux

- *Repiquage de ces plaques dans d'autres plaques (« **plaques d'infection ») :** Ensemencement de 150 µL 2YT+Amp+Tet par puits, avec 10 µL du puits correspondant à la **« plaque référence »**.*
- *Incubation 3 h à 37°C et 500 rpm.*
- *Infection par 6µL de phages helper par puits : incubation 30min sans agitation puis 30 min à 150 rpm.*
- *Ensemencement avec 100 µL à partir des **« plaques d'infection »** dans des **« plaques de culture »** dans 1.5 mL de 2YT+Amp+Kan.*
- *Production des phages o.n. à 30°C et 150 rpm.*
- *Récupération des phages par centrifugation des plaques 1 h à 2000 g.*

✓ Test Elisa

- *On procède comme pour le Test Elisa en pool : Incubation dans chaque puits de la plaque Elisa immobilisée avec la cible avec les phages clonaux correspondant au même puits de la plaque de culture.*
- *La révélation des plaque se fait de la même façon aussi : avec anticorps anti-M13 couplé à la peroxydase.*

Test de reconnaissance cible-*αRep*

✓ Préparation de la plaque Elisa

- *Incubation de la plaque 96 puits o.n. à 4°C avec 100 µL d'une solution de la cible à 20 µg/mL et les colonnes (-) (témoins négatifs) sont incubées avec 100 µL d'une solution de tampon.*
- *3 lavages au TBST.*
- *Blocage 2 h à 4°C et 500 rpm avec 200 µL d'une solution de TBST BSA 3%.*

✓ Préparation des interacteurs

- *Transformation de bactéries Rosetta Blue avec les plasmides des différents interacteurs.*
- *Ensemencement d'une préculture de 10 mL de 2YT+Amp+Glu.*

- *Préparation des cultures à DOi=0.1 : incubation à 37°C, 220 rpm.*
- *Induction de l'expression a DO=0.6 : incubation 4 h.*
- *Récupération des bactéries ont été collectées par centrifugation.*
- *Lyse des culots bactériens : 500 µL de tampon de lyse B-PERII et 3 cycles de congélation-décongélation.*
- *Récupérer les fractions solubles après centrifugation (20 min, 14000 rpm).*

✓ Test d'interaction et révélation

- *Incubation de 200 µL des fractions solubles de chaque interacteur dans la plaque Elisa immobilisée et bloquée : 100 µL pour les puits immobilisés avec la cible et 100 µL pour les puits (-) : 1 h 30 min d'incubation à 4°C.*
- *4 lavages au TBST*
- *Révélation des αRep : Incubation, 1 h à 4°C, avec 100 µL d'une solution de TBST contenant l'anticorps anti-Flag-tag couplé à la peroxydase dilué au 1/20000.*
- *4 lavages au TBST.*
- *Ajout de 100 µL de substrat soluble ont été ajoutés.*
- *Neutralisation de la réaction par ajout de 100 µL de HCl 1 N.*

Résumé

La fonction d'une protéine, qu'il s'agisse de catalyse, de régulation, de transport ou de communication, suppose qu'elle établisse des interactions spécifiques avec un ou plusieurs partenaires moléculaires. Le répertoire des interactions naturellement exercées par des protéines apparaît donc comme très riche par la diversité des objets reconnus. Les anticorps sont de loin les seules protéines permettant de générer des protéines capables d'établir, spécifiquement et à haute affinité des interactions avec diverses cibles. Toutefois, certains inconvénients, liés à leur structure moléculaire, rendent leur production peu efficace et par la suite très coûteuse. Plusieurs approches ont été développées dans le but d'avoir des ossatures protéiques alternatives aux anticorps. Le remodelage de ces ossatures par évolution dirigée a pour but de garder les propriétés d'affinité et de spécificité des anticorps tout en remédiant à leurs inconvénients.

Les travaux décrits dans ce manuscrit concernent le développement d'une nouvelle famille de protéines artificielles alternatives aux anticorps. Le premier chapitre, présente la conception et la construction d'une banque de première génération où les protéines sont formées par la répétition d'un motif *Heat repeat*. En effet, les variants de cette banque dénommée *αRep* ont la même architecture générale mais diffèrent les uns des autres par le nombre de motifs insérés entre les motifs externes et par la séquence dans certaines positions rendues variables au sein de chaque motif (Guellouz et *al.*, 2010). Cette banque nous a permis de valider l'architecture *αRep* choisie. Les protéines de la banque s'expriment sous forme soluble, sont stables et adoptent la structure secondaire attendue.

Cette banque obtenue est loin d'être optimale. Le second chapitre présente alors les approches suivies pour l'amélioration de la qualité de la banque et l'obtention de la banque d'*αRep* de deuxième génération 2.1 contenant $1.7*10^9$ clones indépendants et codants. Les variants ont des séquences aux positions variables qui miment la diversité naturelle et s'expriment sous forme soluble. Cette banque a été par la suite exposée sur phages et des sélections sur des protéines cibles préalablement choisies ont permis d'identifier des interacteurs spécifiques et affins.

Mots clés : *αRep*, protéines artificielles, évolution dirigée, exposition sur phage, reconnaissance moléculaire, structure, interaction protéine - protéine.

Equipe Modélisation et ingénierie des protéines

Institut de Biochimie et de biophysique, moléculaire et cellulaire – UMR8619

Pôle : Ingénierie des protéines Université Paris Sud 11, 91405 Orsay Cedex.