

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
ÉCOLE PRATIQUE DES HAUTES ÉTUDES
Sciences de la Vie et de la Terre
MÉMOIRE

présenté
par

GRUET Antoine

pour l'obtention du diplôme de l'École Pratique des Hautes Études

PROTEINES INTRINSEQUEMENT DESORDONNEES :
DE LA MUTAGENÈSE ALÉATOIRE
À LA RELATION STRUCTURE-FONCTION

soutenu le 14/09/2012
devant le jury suivant :

Président :	Pr Thierry Dupressoir
Tuteur scientifique :	Dr Sonia Longhi
Tuteur pédagogique :	Pr Jean-Michel Verdier
Rapporteur :	Dr Brigitte Meunier-Gontero
Examineur :	Dr Denis Gerlier

Mémoire préparé sous la direction de :

Dr Sonia Longhi

Laboratoire :

Directeur : Dr Yves Bourne

Architecture et Fonction des Macromolécules Biologiques
UMR 7257 CNRS/Université de la Méditerranée
Case 932 - 163 Avenue de Luminy
13288 Marseille cedex 9 (France)

et de

Pr Jean-Michel Verdier

Laboratoire :

Directeur : Pr Jean-Michel Verdier

Mécanismes Moléculaires dans les Démences Neurodégénératives (MMDN)
Université Montpellier 2 - Bât. 24 - Place Eugène Bataillon - CC105
34095 Montpellier cedex 5
EPHE (Sciences de la Vie et de la Terre)

**PROTEINES INTRINSEQUEMENT DESORDONNEES :
DE LA MUTAGENÈSE ALÉATOIRE
À LA RELATION STRUCTURE-FONCTION**

GRUET Antoine

RÉSUMÉ

Les protéines intrinsèquement désordonnées (PID) sont dépourvues de structures secondaire et tertiaire en conditions physiologiques. La plupart des PID sont impliquées dans des fonctions cellulaires intervenant dans des étapes clés de la vie cellulaire, telles que la régulation du cycle cellulaire, la transduction d'un signal et la transcription. Une caractéristique notable des PID est leur capacité à interagir avec de multiples partenaires tout en étant spécifiques. Bien que cette propriété semble liée à leur flexibilité structurale, aucune étude n'a été entreprise jusqu'à présent qui permettrait de comprendre les déterminants moléculaires de la reconnaissance de partenaire.

Le but de ce projet est de tenter de comprendre et de préciser les bases moléculaires déterminant l'affinité et la spécificité des PID lors de la reconnaissance de leur(s) partenaire(s). Pour cela, l'approche *in vitro* utilisée a consisté en la génération de mutants aléatoires au sein de la partie C-terminale de la nucléoprotéine N_{TAIL} (domaine désordonné) du virus de la rougeole. Le domaine N_{TAIL} subit un repliement induit en α -hélice au niveau de la région d'interaction Box2 (aa 489-506), lors de sa liaison avec son partenaire naturel et structuré P_{XD} (partie C-terminale de la Phosphoprotéine). Le système de criblage utilisé est fondé sur le ré-assemblage de la Green Fluorescent Protein (GFP), où N_{TAIL} et P_{XD} sont chacun fusionnés à une moitié de la GFP. Ainsi, l'affinité de l'interaction des variants de N_{TAIL} pour P_{XD}, guide la reconstitution de la GFP.

La mesure de la fluorescence des variants de N_{TAIL}, corrélée à l'analyse de leur séquence peptidique, a permis de confirmer l'implication de certains résidus de la Box2 dans la liaison avec le partenaire. De plus, cette étude met en évidence que la pré-configuration en hélice de la région Box2 facilite l'interaction. Enfin, grâce à cette approche d'évolution descriptive, nous avons identifié des sites de contact primaire, en dehors de la Box2, capables de favoriser l'interaction entre N_{TAIL} et son partenaire P_{XD}.

MOTS-CLÉS: (quatre à huit)

Protéines intrinsèquement désordonnées (PID) ; GFP ré-assemblage ; virus de la rougeole ; interactions protéine-protéine ; évolution dirigée

TABLE DES MATIÈRES

INTRODUCTION

- 1 – Les protéines : généralités
- 2 – Les protéines intrinsèquement désordonnées (PID)
- 3 – Propriétés des protéines intrinsèquement désordonnées (PID)
- 4 – Reconnaissance moléculaire des PID
- 5 – Rôles biologiques
- 6 – Principes généraux de régulation des PID
- 7 - Système modèle
- 8 – Evolution dirigée ou ingénierie des protéines
- 9 – la GFP (Green Fluorescent Protein) : système rapporteur d'interaction
- 10 - Stratégies
- 11 - Vue globale du projet

ABRÉVIATIONS ET ACRONYMES

2YT	Milieu de culture bactérien
AA	Acide aminé
ADN	Acide DésoxyriboNucléique
AKC	Ampicilline + Kanamycine + Chloramphénicol
Ara	Arabinose
CD	Dichroïsme circulaire
C.GFP	Extrémité C-terminale de la GFP
C _{TER}	Extrémité C-terminale
CML	Court Motif Linéaire
DO	Densité Optique
ELM	Eukaryotic Linear Motif
<i>E.coli</i>	<i>Escherichia coli</i>
F	Protéine de fusion
FP	Protéines Fluorescentes
GF	Gel Filtration
GFP	Green Fluorescent Protein
GFPuv	Green Fluorescent Protein ultra violet (variant optimisé pour λ_{ex} =360-400nm)
eGFP	enhanced Green Fluorescent Protein
H	Hémagglutinine
HCA	Analyse des groupes hydrophobes
HeV	Virus Hendra
His tag	Etiquette Histidine
IMAC	Chromatographie d'affinité
IPTG	IsoPropyl- β -D-1-ThioGalactoside
IRF3-RD	Domaine de Régulation de l'Interféron 3
ITC	Titration calorimétrique isothermale
L	Polymérase L
LB	Milieu de culture Lysogeny Broth (bouillon lysogène développé par Bertani)
M	Protéine Matrice
MeV	Virus de la rougeole (measles virus)
ML	Motif Linéaire
MLE	Motif Linéaire Eukaryote
MoRE	Molecular Recognition Element
MoRF	Molecular Recognition Feature
N.GFP	Partie N-terminale de la GFP
N	Nucléoprotéine
NiV	Virus Nipah
N _{TAIL}	Extrémité C-terminale de N
N _{TER}	Extrémité N-terminale
P	Phosphoprotéine

PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PID	Protéine Intrinsèquement Désordonnée
PNT	Extrémité N-terminale de la phosphoprotéine P
RID	Région Intrinsèquement Désordonnée
rpm	Rotation par minute
SAmBA	Simulated Annealing and BAcktracking methods
SCP	Site de Contact Primaire
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
SeV	Virus Sendaï (Sendaï virus)
SplitGFP	GFP coupée en 2 (résidus 157-158)
TB	Milieu de culture Terrific Broth
TFE	2,2,2-TriFluoroEthanol
P _{XD}	Domaine X de P (extrémité C-terminale)
Z	Leucine Zippers (contrôle positif)

INTRODUCTION

1. 1. Les protéines : généralités

Les protéines sont des macromolécules résultant de la polymérisation d'acides aminés. Ces molécules sont composées essentiellement de carbone C, hydrogène H, oxygène O, azote N et souvent les éléments suivants sont présents : le soufre S et le phosphore P. Les acides aminés sont liés entre eux par une liaison amide particulière : la liaison peptidique.

Lors de la traduction, les acides aminés s'enchaînent dans un ordre dicté par la séquence nucléotidique du gène. C'est sous cette forme d'enchaînement que l'on définit la structure primaire d'une protéine. D'autres niveaux d'organisation existent permettant à la protéine d'adopter une organisation tridimensionnelle qui lui est propre. Ce processus d'arrangement tridimensionnel est appelé repliement d'une protéine. Il a été proposé une description hiérarchique des différents niveaux de structuration d'une protéine, allant de la séquence en acides aminés à l'arrangement spatial des sous-unités. Il existe trois niveaux de structuration supplémentaires à la structure primaire :

- a. La structure secondaire définie par la présence d'éléments tels que hélice alpha, brin beta, coude ou boucle.
- b. La structure tertiaire, définie par le repliement de la protéine dans l'espace, comprenant les éléments de la structure secondaire. Ainsi des résidus éloignés dans la structure primaire peuvent se retrouver à proximité (ex : triade catalytique).
- c. La structure quaternaire définie par l'association de plusieurs sous-unités ou chaîne polypeptidique structurées au niveau tertiaire. Ces sous-unités s'associent de façon stoechiométrique grâce à des liaisons non covalentes.

Ces niveaux de structuration sont essentiels à la fonction de la protéine. Si une protéine perd sa structure tridimensionnelle stable et unique, en raison de facteurs physiques ou chimiques, la protéine est dite dénaturée et perd son activité (ex : insuline). Un retour à l'état natif (initial) avec conservation des propriétés physico-chimiques de la protéine est possible dans certains cas (exemple de la protéine kinase ou de la ribonucléase [5, 6] [7]). La structure tri-dimensionnelle est donc essentielle à la fonction, [8] ce qui conduit au paradigme structure-fonction. Ce dernier postule que la fonction d'une protéine dépend de la structure et que la connaissance détaillée de celle-ci fournit des informations sur sa fonction. L'emplacement précis dans l'espace d'un ou de plusieurs résidu(s) crée un environnement physico-chimique particulier nécessaire à divers événements, comme la fixation de ligands, la formation de sites catalytiques, l'assemblage de complexes macromoléculaires ou la translocation d'ions ou de petites molécules. Ce paradigme est éminemment étayé par les nombreuses structures déposées dans la Protein Data Bank (PDB), détaillant par exemple le mécanisme d'action d'enzymes, de récepteurs, de transporteurs, de canaux membranaires. Dans le cas des enzymes, le site actif est composé par une combinaison d'acides aminés (portant l'activité catalytique), qui sont parfaitement positionnés grâce à une structure 3D précise et stable. Bien que la fonction soit intimement corrélée à une structure unique et stable, de nombreuses preuves expérimentales montrent que le repliement sous une forme 3D unique et stable n'est pas obligatoire pour obtenir des molécules biologiquement actives. Depuis quelques années, le paradigme structure-fonction a été remis en question, pour être finalement étendu. En effet, un nombre croissant d'études mettent en évidence que l'absence de structure 3D précise, unique et stable chez certaines protéines ne signifie pas nécessairement une absence d'activité en conditions physiologiques. Ces protéines, dites nativement non-structurées, ou partiellement repliées, ou intrinsèquement

désordonnées, constituent une nouvelle définition de la relation structure-fonction [9].

Les principes physiques gouvernant l'organisation de ces deux classes de protéines (ordonnées et désordonnées) sont identiques. L'existence des protéines désordonnées fonctionnelles a donc conduit à étendre le paradigme puisque la fonction d'une protéine reste dépendante de sa structure. Dans ce cas, l'absence de structure rigide confère aux protéines désordonnées leurs propriétés et fonctions.

a. 2. Les protéines intrinsèquement désordonnées

Une protéine intrinsèquement désordonnée (ou une région intrinsèquement désordonnée) est une protéine (ou une région) biologiquement active existant comme un ensemble de conformères dynamiques (essentiellement dans un état étendu), avec peu ou pas de structure secondaire stable en conditions physiologiques. La figure 1, ci-après, indique les différents niveaux de structuration existant entre un état complètement ordonné et un état totalement désordonné [7] à [12].

Depuis les années 90, on observe dans la littérature une augmentation du nombre d'études portant sur ces protéines.

Plus d'un tiers des protéines eucaryotes contiennent des régions désordonnées de longueur supérieure à 30 résidus ce qui souligne leur importance physiologique [10]. Leur composition en acides aminés montre un enrichissement en acides aminés induisant le désordre et un appauvrissement en acides aminés induisant la structuration [11]. En effet, l'analyse de la séquence de 275 protéines naturellement repliées et de 91 protéines naturellement non-repliées a permis de déterminer qu'un appauvrissement en résidus hydrophobes et un enrichissement en résidus chargés, étaient un pré-requis pour l'obtention d'une structure relâchée, non compacte, au sein d'une protéine. Plus précisément, certains résidus favoriseraient le désordre, tels que Ala, Arg, Gly, Gln, Ser, Glu, Lys et Pro, alors que d'autres favoriseraient la structuration de la séquence : Ile, Leu, Val, Trp, Tyr, Phe, Cys et Asn [12]. Grâce à ce biais compositionnel, des logiciels de prédiction de désordre ont vu le jour : PONDR, GlobProt, DisEMBL, Disopred2, RONN, DISpro...

Au sein de l'équipe, un métaserveur pour la prédiction de désordre a été créé et développé : MeDor [13]. Ce métaserveur permet d'interroger simultanément plusieurs logiciels de prédiction en réseau, afin d'afficher leurs résultats.

A partir d'une séquence protéique, une représentation graphique permet de visualiser les régions désordonnées ainsi que les structures secondaires. Le logiciel comporte :

1. Un module de prédiction de structure secondaire
2. L'affichage de la séquence peptidique sur une ligne unique
3. Une numérotation des résidus
4. Un tracé HCA (analyse des groupes hydrophobes) [14], permettant de distinguer les régions globulaires (riches en amas hydrophobes) des régions non globulaires (pauvres en amas hydrophobes) et de visualiser les structures secondaires parmi les régions globulaires
5. Le résultat de plusieurs prédicteurs de désordre concernant l'étendue de la (des) région(s) désordonnée(s) le long de la séquence étudiée
6. Un prédicteur (Phoebius) de domaines transmembranaires et de peptides signaux

a.

b. 3. Propriétés des protéines intrinsèquement désordonnées (PID)

a. 3.1 État en solution

b.

Les protéines intrinsèquement désordonnées (PID) ne possèdent pas de structure tridimensionnelle unique en solution. Des études conformationnelles et spectroscopiques ont montré que les PID ne constituent pas une famille homogène, mais existent en tant qu'ensembles de conformères [15] à [16] :

Ces protéines, en général, peuvent exister sous 4 états conformationnels différents, soit un degré de compaction décroissant :

- Ordered (O) = ordonnée
- Molten globule (MG) = forme structurée relâchée
- Pre Molten Globule (PMG) = forme structurée étendue
- Coil-like (Coil) = forme étendue désordonnée

•

a. 3.2 Flexibilité structurale des PID

b.

La flexibilité conformationnelle des protéines désordonnées leur confère de nombreux avantages fonctionnels, tels qu'une forte tolérance à la promiscuité, une spécificité de reconnaissance combinée à une faible affinité et une augmentation des vitesses d'interaction [17]. Ces avantages fonctionnels confèrent des propriétés essentielles dans les interactions protéine-protéine telles que celles impliquées dans des processus de signalisation et de régulation (exemples : les protéines p27 et p53) [18]. Un exemple extrême illustrant cette importance est la protéine ribosomale YP2b qui possède toutes les caractéristiques d'un état molten globule (MG) en l'absence d'ARN (Zurdo et al., 1997). La grande plasticité liée à sa structure joue un rôle majeur dans l'assemblage des différentes structures composant le complexe ribonucléoprotéique qu'est le ribosome. Le repliement induit par les protéines et l'ARN, dirige l'assemblage des multiples composants dans l'ordre adéquat aboutissant à la structure finale fonctionnelle [19]. L'absence de structure 3D rigide et l'exposition de régions clés au solvant, fournissent un mécanisme simple de régulation par phosphorylation ou par l'intermédiaire d'une interaction avec d'autres composants de la machinerie cellulaire [19]. La flexibilité structurale des PID permet aussi une augmentation de la surface d'interaction par rapport à celle des protéines globulaires de même taille.

Mais ces particularités structurales entraînent aussi certains inconvénients. La flexibilité due à l'état étendu de la séquence peptidique entraîne une plus grande sensibilité aux protéases [20].

a. 4. Reconnaissance moléculaire des PID

Le fonctionnement d'une physiologie bien réglée de tout organisme vivant, est fondé sur un ensemble d'interactions protéiques coordonnées : l'interactome. Chez l'Homme, la taille de l'interactome serait estimé à 650 000 interactions protéine-protéine [21]. Le contrôle de ces interactions coordonnées repose sur la reconnaissance de régions spécifiques d'identification souvent localisées au sein de régions intrinsèquement désordonnées [12]. Un autre mécanisme de liaison spécifique des PID est caractérisé par les "sites de contact primaire" (SCP). Ces SCP sont de courts motifs de reconnaissance localisés dans des régions désordonnées, plus exposées que d'autres régions. Ils peuvent donc servir de premier site de contact avec leur partenaire et sont donc cinétiquement séparés de la liaison de la région d'interaction avec le partenaire [22]. Certains éléments de reconnaissance sont souvent détectés comme de courts motifs bien conservés, nommés aussi séquences consensus. La généralisation de cette relation a conduit au concept de Motif Linéaire (ML), aussi noté Motif Linéaire Eucaryote (MLE), ainsi que Court Motif Linéaire (CML). Il a été montré que ces ML sont souvent localisés dans des régions désordonnées [23] à [24].

a. 4.1 Repliement induit

Certains complexes cristallisés comportent une longue séquence peptidique (le partenaire) et une séquence peptidique entre 10 et 70 résidus, plus petite. Cette petite séquence peptidique (de la PID) contient les résidus qui créent la liaison avec le partenaire, et est appelée MoRE (Molecular Recognition Element) ou MoRF (Molecular Recognition Feature). Cette petite région nativement désordonnée est capable de se structurer en présence du partenaire : ce phénomène est appelé repliement induit. Ainsi, les résidus du MoRE impliqués dans l'interaction sont ajustés à la structure du partenaire. Les MoRE se classent en 4 catégories structurales dépendantes de la structure secondaire dominante, selon leur état de liaison avec leur partenaire : α -MoRE (hélice alpha), β -MoRE (brin bêta), i-MoRE (forme irrégulière), et mixed-MoRE (multiple). La majorité des PID subissent un changement structural en passant d'un état désordonné vers un état ordonné lors de l'interaction avec leur partenaire [1], [11], [19], [25]. Lorsque des régions désordonnées se lient à un partenaire, la transition désordre-ordre comporte une pénalité entropique. Par conséquent, l'affinité, qui est fonction de la variation d'énergie libre, est moindre. Ce phénomène caractérise alors la nature des interactions des régions désordonnées, à savoir une liaison très spécifique (la PID reconnaît précisément ce partenaire) et une faible affinité (l'interaction est réversible) [11]. Il existe des exceptions où l'affinité pour le partenaire est élevée (exemple : la partie C-terminale de la nucléoprotéine du virus de la rougeole). La plupart des PID recrutent des partenaires différents. Ainsi, une interaction définie par une haute spécificité et faible affinité est une combinaison idéale pour assurer la réversibilité d'une interaction, permettant une médiation du message de signalisation contrôlé au sein de la cellule. De plus, la capacité d'une PID à interagir avec plusieurs partenaires conduit à des effets biologiques multiples, ce qui comporte un avantage en termes de compaction génétique.

Il a aussi été observé certaines exceptions au modèle d'interaction « clé-serrure ». En effet, le site actif n'est pas, de façon exceptionnelle, entièrement complémentaire au substrat. Le modèle ne peut expliquer complètement le mode d'interaction des enzymes avec leur substrat, et ne peut décrire exhaustivement la stabilisation de l'état de transition nécessaire lors de l'interaction. Koshland [26] a proposé une modification du modèle d'interaction « clé-serrure ». Il suggère que le substrat ne se lierait pas à un site actif rigide, mais plutôt à un site actif possédant une structure adaptée lui permettant de réaliser la fonction catalytique requise. Le site actif posséderait une flexibilité nécessaire pour atteindre le plus haut état d'énergie intermédiaire de conversion (état de transition) lors d'une réaction enzymatique. Il s'agirait d'une adaptabilité induite (induced fit).

Les PID représentent un cas extrême d'adaptabilité induite, elles peuvent donc facilement interagir en changeant de conformation lors de l'interaction avec différents partenaires [12]. Plusieurs études ont mis en évidence que les PID possèdent une structure résiduelle, correspondant à une préférence conformationnelle pour les structures qu'elles adoptent dans leur forme complexée. Par exemple, la protéine p53 [27] possède des structures résiduelles α -hélicales et se replie en hélice- α lors de la liaison avec des partenaires. Certains complexes issus du repliement induit ont pu être cristallisés, du fait d'une importante transition désordre-ordre, donnant ainsi des formes uniques et stables. Cependant l'état désordonné peut persister même après interaction avec le partenaire, un phénomène nommé « fuzziness ». Ainsi, des régions de taille plus ou moins importante, peuvent maintenir un état désordonné au sein du complexe.

La persistance d'un état désordonné dans les régions encadrant les MoRE (régions d'interaction structurées) au sein de la forme complexée, a été nommée [29] flexibilité en bordure, en opposition à la flexibilité aléatoire, où les PID demeurent entièrement désordonnées lors de l'interaction avec leur partenaire. Ces observations suggèrent que le repliement après l'interaction n'est pas nécessaire pour maintenir la reconnaissance moléculaire et la fonction de signalisation des PID [30].

a. 4.2 Capacités de liaisons et promiscuité

Les PID sont capables d'interagir avec de multiples partenaires. Elles servent souvent de protéines « pivot » dans les réseaux d'interaction protéiques. Ces protéines « pivot » sont cruciales pour les organismes vivants et leur délétion au sein d'un organisme peut s'avérer létale. Des exemples de PID « pivot » possédant des dizaines, voire des centaines, de partenaires de liaison, sont les protéines : p53, α -synucléine et BRCA1.

a. 4.3 Vitesse de réaction - Association/Dissociation

L'absence de structure 3D unique et stable, confère de nombreux avantages. Parmi ces derniers, notons une vitesse de liaison élevée, ainsi qu'un volume conformationnel important. Ces deux propriétés sont à la base du « fly-casting ». Un mécanisme de liaison où la protéine non repliée se lie d'abord faiblement à un point d'ancrage sur le partenaire et à une distance relativement importante du site de liaison. Puis la protéine désordonnée approche du site de liaison au fur et à mesure que le MoRE se replie, pour enfin adopter la conformation finale et disposer correctement les résidus impliqués dans l'interaction [31].

5. Rôles biologiques

Les PID réalisent d'importantes fonctions au sein de l'organisme et sont impliquées dans de nombreux processus vitaux. Une grande flexibilité, leur permettant d'adopter plusieurs conformations, est fondamentale pour assurer leurs rôles biologiques. Les protéines ayant des fonctions de signalisation, de régulation ou de liaison, telles que les facteurs de transcription et les protéines de signalisation, sont enrichies en région intrinsèquement désordonnées [10]. Les nombreux exemples de domaines non structurés en solution devenant structurés lors de la liaison avec le partenaire ont été identifiés comme faisant partie des protéines contrôlant le cycle cellulaire, ainsi que celles régulant la transcription et la traduction [19]. Du fait de leur rôle central dans des fonctions biologiques importantes, des dysfonctionnements affectant ces protéines conduisent à des états pathologiques. En effet, en combinant les résultats des prédicteurs de désordre et les informations reliant les protéines mises en cause lors d'une maladie spécifique, il a été montré que les protéines impliquées dans le cancer, les maladies cardio-vasculaires, les maladies neurodégénératives, les amyloïdoses et plusieurs autres maladies [12] présentaient un haut niveau de désordre. Par exemple, la protéine Tau, en subissant une transition de type bêta, conduit à la formation d'agrégats au niveau neuronal, responsables de la maladie d'Alzheimer et autres tautopathies [32]. Les régions désordonnées peuvent aussi avoir une autre fonctionnalité. Par exemple, Nudix, une enzyme (EC=3.6.1.-) provenant de l'organisme *Deinococcus radiodurans* (résistant aux radiations et à la dessiccation), corrige certaines altérations localisées au niveau de l'ADN, et aurait un rôle prépondérant durant la phase de récupération après une exposition à des radiations ionisantes. Cette enzyme utilise le magnésium comme co-facteur lors des réactions d'hydrolyse des nucléotides triphosphate. D'un point de vue structural, Nudix possède à ses extrémités NTER et CTER des régions désordonnées, lui permettant d'augmenter sa probabilité de localisation dans les poches d'eau résiduelles, maintenant ainsi l'enzyme suffisamment hydratée pour être fonctionnelle. Ainsi, par extension, les RID (Régions Intrinsèquement Désordonnées) pourraient augmenter l'hydrophilicité des parties hydrophobiques fonctionnelles et influencer leur localisation cellulaire [33].

6. Principes généraux de régulation des PID

En raison de leurs importantes propriétés fonctionnelles et de leurs rôles essentiels dans le métabolisme cellulaire, la présence des PID a besoin d'être contrôlée de manière fine. Les PID sont régulées grâce à plusieurs mécanismes agissant aux différentes étapes de leur synthèse. Les mécanismes de régulation transcriptionnelle, traductionnelle et post-traductionnelle contrôlent la disponibilité de ces protéines, ainsi présentes en quantité contrôlée et durant un laps de temps limité. Les procédés de régulation influent sur i) les niveaux d'expression géniques, ii) le taux de dégradation des ARN messagers, iii) le ciblage des micro-ARN (miRNA), répresseur post-transcriptionnel et iv) l'ubiquitination, un indicateur cellulaire de protéolyse [34]. La régulation sur de courtes périodes est principalement due à la régulation post-transcriptionnelle via une diminution de la demi-vie des transcrits et à une protéolyse accrue. En effet, les analyses informatiques montrent que les transcrits codant pour ces protéines sont enrichis en sites de fixation aux miRNA et ont des temps de demi-vie plus courts. De plus, par comparaison aux protéines ordonnées, les PID sont enrichies en sites d'ubiquitination informatiquement prédits. Ainsi, les PID et/ou leurs transcrits possèdent des

propriétés qui sont des indicateurs d'une plus faible expression génique. Les domaines non structurés sont donc présents chez les protéines qui sont ciblées pour être dégradées rapidement [19].

Edwards et al. [34] suggèrent que l'enrichissement des PID en sites prédits comme fixant les miRNA et en sites d'ubiquitination empêcherait l'accumulation des PID, ce qui serait délétère pour la cellule. Toutefois, cette même étude identifie un groupe de PID qui sont hautement exprimées, suggérant que certaines PID bien identifiées dans certains tissus sont nécessaires à un niveau élevé, pour le bon fonctionnement du métabolisme cellulaire [34]. Vavouri et al. [35] ont cherché à comprendre pourquoi certains gènes étaient nocifs lors d'une sur-expression dans une cellule. Ces gènes sont dits dosage-dépendants, et codent pour des PID. En étudiant une grande quantité de données issues des études concernant la levure *S.cerevisiae*, ces auteurs ont conclu que les PID ont un important déterminisme de toxicité dosage-dépendant. Plus particulièrement, ils ont observé que les protéines contenant des motifs linéaires peptidiques sont particulièrement dosage-dépendants. De plus, ces auteurs ont démontré que ce type de protéine est très finement régulée à la fois au niveau des ARN messagers (mRNA) et au niveau protéique, et suggèrent que ce contrôle agit pour limiter les augmentations potentiellement néfastes de la concentration en PID. Que ce soit chez le nématode, la drosophile ou chez l'Homme, les mêmes propriétés relient les gènes dosage-dépendants et les oncogènes. L'explication avancée est que les régions désordonnées sont susceptibles d'établir des interactions moléculaires aspécifiques via les motifs linéaires, ce qui est la cause de leur toxicité lors de leur sur-expression. Cependant, toutes les PID ne sont pas toxiques lors d'une sur-expression, comme par exemple les facteurs de transcription, dont la sur-expression a été montrée expérimentalement comme bénigne chez la levure [10]. Les PID peuvent aussi être l'objet de diverses modifications post-traductionnelles facilitant la régulation de leurs fonctions et de leur temps de demi-vie dans la cellule (ex : phosphorylation) [10]. De plus, les interactions impliquant des motifs linéaires sont très répandues dans la nature, et les domaines protéiques liant les motifs linéaires sont présents dans de nombreuses familles de protéines, avec des préférences de liaison de séquences similaires [10].

Pour résumer, une PID est caractérisée par [20]:

- _ Une composition spécifique en acides aminés comprenant peu de résidus hydrophobes et une charge totale nette élevée ;
- _ Des propriétés hydrodynamiques typiques d'un état « random coil » ou « Pre Molten Globule », à savoir un état hydrodynamique relâché ;
- _ Une faible quantité de structures secondaire et tertiaire ;
- _ L'absence de noyau structuré de façon très compacte ;
- _ Une flexibilité conformationnelle élevée ;
- _ L'adoption d'une structure relativement rigide induite par la présence du partenaire naturel.

Dans ce projet, le modèle d'étude de PID utilisé est un domaine d'une protéine du virus de la rougeole, appartenant au complexe répliquatif. Il s'agit de la région C-terminale de la nucléoprotéine (N) appelée NTAIL dont le partenaire naturel est le domaine X de la phosphoprotéine (PXD) [36]. A ce jour, aucune étude n'a permis de relier les caractéristiques moléculaires, responsables de la sélectivité des PID, à leurs propriétés de liaison. Le partenariat moléculaire NTAIL-XD servira de modèle d'étude en vue d'identifier les bases moléculaires de l'affinité et de la spécificité lors de la reconnaissance de partenaires par les PID.

7. Système modèle

Le système modèle utilisé dans cette étude concerne deux protéines virales. Le virus de la rougeole appartient à l'ordre des *Mononegavirales*, à la famille *Paramixoviridae* et au genre *Mobilivirus*. Le virus est enveloppé d'une double couche lipidique, exposant à sa surface des protéines. Le génôme viral est un ARN simple brin de polarité négative, non segmenté et encapsidé hélicoïdalement. Le génome code pour 6 protéines : N (Nucléoprotéine), P (Phosphoprotéine), M (Matrice), F (Fusion), H (Hémagglutinine), L (Large polymerase).

La polarité négative implique que l'ARN ne peut pas être traduit directement par la cellule hôte. Le brin négatif nécessite une étape de transcription afin d'obtenir des ARN messagers. Lors de la réplication, une copie complémentaire du génome entier est synthétisée. Le génome viral (brin – ou brin +), est enveloppé par la protéine N pour former la nucléocapside. Cet ensemble ARN + capsid sert de substrat à la polymérase L lors de la transcription (1) et de la réplication (2).

La nucléoprotéine (N) permet de recruter la polymérase (L) via son interaction avec la phosphoprotéine (P). Sans la P, la L ne peut plus être recrutée et le virus ne peut plus réaliser les étapes de transcription et de réplication. L'interaction entre N et P se fait grâce à leurs domaines C-terminaux : N_{TAIL} et P_{XD} respectivement.

La nucléoprotéine (N) du virus de la rougeole est composée de 525 résidus et s'organise en 2 parties :

- une partie ordonnée appelée core : Ncore, aa 1 - 400
- une partie désordonnée appelée queue : NTAIL, aa 401 - 525.

La région C-terminale de la protéine virale N (NTAIL) est le centre du projet. NTAIL est une région désordonnée. Il a été montré que cette région était capable de lier la partie C-terminale de P, nommée le domaine X (PXD), grâce à une région particulière appelée Box2. Localisée au sein de NTAIL, cette région a une composition enrichie en acides aminés hydrophobes, qui contraste avec le reste de la séquence. De plus, la région Box2 subit un repliement induit en hélice _ lors de la liaison avec le partenaire naturel PXD [37].

La région Box2 s'étend des résidus R489 à E506. La prédiction d'une structure _-hélicale comprend les résidus S488 à G503. La région d'interaction liant PXD s'étend des résidus 457 à 507. [2].

Les données RMN de Gely et al. [38] indiquent que les résidus impliqués dans l'interaction vont de R490 à I504. La protéine virale NTAIL servira de modèle d'étude, avec comme partenaire d'interaction PXD. Chez le virus de la rougeole, la phosphoprotéine (P) est composée de 507 résidus et l'extrémité C-terminale (appelée domaine X ou PXD ; aa 458 – 507) s'organise en un domaine structuré de 49 résidus et adopte un repliement en 3 hélices alpha.

Afin de mettre en évidence l'interaction entre Ntail et PXD, nous avons utilisé une méthode de révélation fondée sur la fluorescence. Le signal de fluorescence est émis par la Green Fluorescent Protein (GFP). Les différences d'intensité de fluorescence sont corrélées à une modification de l'interaction entre les deux protéines. Le projet s'appuie sur l'évolution dirigée de la protéine NTAIL par mutagenèse aléatoire. Cete dernière est réalisée sur la séquence codant pour la protéine NTAIL entière afin d'appréhender, sans a priori, l'impact des substitutions sur l'interaction. Les informations de toutes les études connues à ce jour ne concernent que quelques parties de NTAIL.

J'ai généré de façon aléatoire des mutations au sein de Ntail afin de produire une banque de mutants. J'ai ainsi créé des variants aléatoires de Ntail et j'ai évalué l'impact des substitutions sur sa capacité à interagir avec son partenaire naturel PXD par l'intermédiaire de la protéine rapporteur. Après séquençage des mutants sélectionnés, j'ai réalisé l'analyse des résultats afin de dégager des axes de recherche permettant d'élargir les connaissances concernant le mécanisme d'interaction de Ntail et d'en déduire les bases moléculaires de la reconnaissance de partenaires.

8. Évolution dirigée ou ingénierie des protéines

En mimant l'évolution naturelle, l'évolution dirigée vise à générer une protéine possédant une nouvelle activité en criblant ou en sélectionnant la fonction désirée, depuis un vaste ensemble de variants de cette protéine [3] -[39].

a. 8.1 Principes

L'évolution dirigée comprend deux étapes. 1) La constitution d'une banque de mutants, où la diversité génétique est créée, et 2) la sélection ou le criblage des variants, en fonction de certaines propriétés modifiées que l'on recherche [40]. L'évolution dirigée consiste en la création d'un ensemble de mutants d'un gène, constituant la banque de mutants. Lorsque ces gènes mutés sont exprimés chez un organisme hétérologue (E.coli dans notre projet), on obtient des variants de la protéine à étudier. Les variants fonctionnels sont ensuite évalués lors du test de criblage, de manière à identifier ceux possédant les propriétés attendues. Les gènes mutés les plus intéressants (gènes codants pour les protéines ayant les propriétés recherchées) peuvent être soumis à un nouveau cycle de mutagenèse dans le but d'accumuler les mutations et d'accentuer l'(es) effet(s) recherché(s). Cette évolution, impliquant un ou plusieurs cycles, est dépendante des objectifs fixés ainsi que des effets observés. Plusieurs propriétés des protéines peuvent être sujettes à l'évolution dirigée, telles qu'une thermostabilité accrue, une augmentation des performances catalytiques ou l'acquisition de nouvelles propriétés (vis à vis d'un substrat), l'amélioration de l'activité d'une protéine d'un organisme mésophile à faible température, ou bien une conservation de l'efficacité catalytique de l'activité d'une enzyme en présence de solvants différents du milieu originel (solvants organiques, mélange aqueux-solvant organique) [41]. La plupart des méthodes d'ingénierie de protéines reposent sur l'introduction de substitutions au sein de protéines d'origine naturelle qui possèdent déjà une partie des propriétés désirées. L'évolution dirigée a démontré maintes fois son efficacité, et se présente comme une alternative aux approches rationnelles de l'ingénierie des protéines [42]. Ceci restera le cas en l'absence d'une compréhension complète de la façon dont les acides aminés dictent la fonction biologique [43]. Les pré-requis nécessaires pour une évolution dirigée réussie sont : l'expression d'une protéine fonctionnelle dans un système d'expression adapté, une méthode de criblage rapide et efficace quant aux caractéristiques recherchées, ainsi qu'une stratégie bien construite pour explorer les différentes possibilités du paysage protéique. Le défi majeur consiste à identifier une voie d'évolution réalisable, qui résultera en une augmentation des caractéristiques désirées [42].

a. 8.2 Techniques

Le développement des méthodes optimisées pour l'introduction aléatoire de mutations au sein d'une séquence nucléotidique, a facilité les premières expériences d'évolution dirigée. Ainsi, les mutations sont régulièrement introduites le long d'une séquence d'ADN. Wong et al. [40] ont développé un « Programme Assistant la Mutagenèse » (MAP) qui est une méthode d'analyse statistique, fournissant aux ingénieurs un outil performant pour définir des stratégies d'évolution dirigée efficaces. Pour aboutir à ce programme, ils ont comparé 19 méthodes de mutagenèse appliquées à 4 gènes codant pour des protéines différentes [40]. L'évaluation de l'efficacité des méthodes repose sur trois indicateurs de biais prenant en compte les substitutions générées au niveau protéique : i) indicateur de structure des protéines, ii) indicateur de diversité amino-acidique couplé à un coefficient de diversité des codons et iii) indicateur de diversité chimique.

Pour l'introduction aléatoire de mutations, plusieurs méthodes peuvent être utilisées seule ou en combinaison. Elles sont fondées sur l'utilisation [41], [44] :

- _ d'enzymes introduisant des erreurs (error-prone PCR ou epPCR), suivies ou non de mélange aléatoire de fragments d'ADN portant des mutations (DNA shuffling) ;
- _ d'agents chimiques mutagéniques (hydrazine, méthanesulfonate d'éthyle EMS ...) ;
- _ ou de souches mutagéniques (Mutator strain E.coli).

Ces méthodes de création de mutants nécessitent une méthode de récupération efficace, indispensable à la détection de nouveaux variants protéiques. Il existe deux méthodes permettant de choisir les variants parmi une banque : la sélection et le criblage.

a. 8.3 Sélection ou criblage ?

La sélection consiste à obtenir uniquement les mutants désirés. On suppose donc que le phénotype (survie) est lié aux caractéristiques génétiques qui nous intéressent. Le criblage consiste à étudier les clones disponibles (tous ou un nombre limité), et à tous les tester selon le critère d'évolution recherché. Si le test est positif, le mutant est conservé. Les méthodes de criblage ou de sélection peuvent être fondées sur 1) un phénotype facilement exploitable :

- par criblage grâce à une couleur : verte pour la GFP, rouge pour une protéine liant l'hème [45], ou la couleur bleue de la plastocyanine [46].

- par criblage grâce à une réaction enzymatique où l'un des produits de la réaction peut être dosé [47] .

- par criblage visuel fondé sur les propriétés hydrolytiques d'enzymes sécrétables lors d'une croissance sur milieu adapté [41], ou fondées sur 2) un criblage automatisé en plaque 96 puits. L'automatisation permet une accumulation de données et une facilité d'analyse appréciables lors de la manipulation d'une banque de mutants [48]. Par exemple, certains auteurs ont utilisé un screening automatisé et des substrats chromogéniques pour détecter les caractéristiques recherchées au sein de nouvelles enzymes créées par mutagenèse aléatoire [49], ou enfin sur 3) un tri de cellules. Dans une optique de criblage automatisé à haut débit plus performant, il est indispensable d'améliorer la miniaturisation des volumes. Des systèmes de prélèvements ont été développés et sont capables de gérer des volumes de l'ordre du picolitre (pL) à une vitesse allant jusqu'à 10 000 gouttes/seconde [48], [50]. La miniaturisation, ainsi que l'automatisation, améliorent les capacités de criblage du nombre de variants et permettent donc d'accroître la puissance de l'évolution in vitro, en identifiant des événements rares. D'autres méthodes de criblage plus sophistiquées peuvent aussi exploiter une attraction chémo-tactique [51], ou des réactions immunologiques [52], [53]. Les méthodes de sélection peuvent être fondées sur le ribosome display et le RNA display [54]. Ces méthodes servent à réaliser une évolution dirigée de protéines, dans le but de créer des protéines capables de fixer un ligand déterminé. Les protéines en cours de traduction, associées à leur ARNm, se lient à un ligand immobilisé choisi. Le complexe ARNm-protéine ainsi fixé est ensuite retro-transcrit en ADN puis amplifié par PCR. L'ADN récupéré code ainsi pour une protéine aux propriétés nouvelles, ie capable de fixer le ligand désiré.

9. La GFP (Green Fluorescent Protein) : système rapporteur d'interaction

Les protéines produisant de la lumière impliquent deux protéines physiquement proches l'une de l'autre. La première protéine est productrice de lumière (par exemple, l'aequorine lors de la présence d'ions calcium). L'énergie libérée permet à la seconde d'émettre une lumière de longueur d'onde supérieure à la première (décalée vers le rouge). Le pic d'absorption de la deuxième protéine correspond au pic d'émission de la première protéine.

La GFP est l'une des protéines de la seconde catégorie. Chez la méduse *Aequora victoria*, la chimi-luminescence bleue émise par l'aequorine est convertie en vert par la GFP, probablement pour réduire la dispersion de la luminosité et permettre une meilleure diffusion dans l'océan [55]. La Green Fluorescent Protein (GFP) est une protéine composée de 238 acides aminés et possède une masse moléculaire d'environ 27 kDa. La longueur d'onde d'excitation maximale se situe à 375 nm (longueur d'onde correspondant à la couleur bleue) et la longueur d'onde d'émission maximale se situe à 508 nm (longueur d'onde correspondant à la couleur vert).

La GFP se structure en forme de canette, où les parois sont constituées par 11 brins beta, le centre actif responsable de la fluorescence se plaçant au centre. La formation du centre actif ne requiert aucun co-facteur propre à la méduse (organisme originel), mais nécessite cependant une phase de maturation pour être fonctionnel. Cette phase de maturation, auto-catalytique, implique les résidus Ser65, Tyr66, Gly67 de la GFP, ainsi que la présence d'oxygène. La maturation aboutit à la structure suivante : 4-(p-hydroxybenzylidene)-imidazolidin-5-one. La mutagenèse combinatoire a mis en évidence que le résidu Gly67 est indispensable à la formation du chromophore [55].

L'enchaînement de ces trois résidus se retrouve dans d'autres protéines sans pour autant conduire à la formation d'un fluorophore. Le mécanisme séquentiel de la formation du chromophore est initié par une cyclisation rapide entre les résidus Ser65 et Gly67, pour former un intermédiaire (imidazolin-5-one), qui est suivie d'une étape d'oxydation, limitante et très lente (plusieurs heures), de la chaîne latérale de la Tyr66 par l'oxygène [55].

Deux conditions semblent être nécessaires pour cette formation : la proximité stérique des résidus 65 et 66 (Ser et Tyr) et la présence d'un couple acide/base pour catalyser la cyclisation [56]. La comparaison de plus de 100 séquences révèlent 4 résidus très conservés. Il s'agit des résidus Tyr66, Gly67 du tripeptide chromophore et de l'Arg96 et Glu222. Le résidu Gly67 est le nucléophile de la réaction initiale de la formation du chromophore (Ser 65 et Tyr66 appartenant physiquement au chromophore), tandis que les résidus Arg96 et Glu222 sont les résidus catalytiques [57]. La formation du chromophore donnerait une fluorescence de moins en moins abondante lorsque la température est supérieure à 30°C. Une fois formée, la protéine GFP est très stable.

Il est possible de classer les fluoro-protéines (FP) en deux groupes, sur la base de leur procédé de maturation du chromophore [57]. Les protéines fluorescentes de la couleur allant du cyan au vert ont un chromophore identique à la GFP, tandis que les FP de couleur jaune à rouge subissent une réaction d'oxydation additionnelle, comme par exemple la DsRed. Cette réaction étend physiquement le chromophore qui peut ainsi émettre à des longueurs d'ondes plus grandes [58], [59], [60].

Dans ce projet, la GFP sert de rapporteur d'interaction entre deux protéines partenaires : NTAIL et PXD. L'interaction entre ces deux protéines est révélée grâce au ré-assemblage des deux fragments de la split-GFP (GFP coupée en 2 au niveau des résidus 157 et 158) en une GFP fonctionnelle [61]. En effet, Le fragment N-terminal de la split-GFP est lié à Ntail, tandis que le fragment C-terminal est lié à PXD. La reconstitution de la GFP, donc sa capacité à fluorescer, est uniquement conditionnée par l'interaction des deux protéines fusionnées. Lorsque la GFP est reconstituée, le chromophore est maintenu fonctionnel, en dépit d'une éventuelle dissociation des deux partenaires. Ceci conduit donc à une accumulation de protéines fonctionnelles fluorescentes dans le temps. Plus précisément, l'apparition de la fluorescence est aussi corrélée à l'affinité de l'interaction des deux partenaires fusionnés. La coupure de la protéine se situe au niveau des résidus 157-158, au milieu d'une structure flexible [62]. La séquence du chromophore (résidus 65, 66 et 67) est intact. Deux GFP ont été utilisées au cours de ce projet: la GFPuv et la eGFP (enhanced GFP). La eGFP possède une fluorescence nettement supérieure à celle de la GFPuv du fait de quelques mutations intégrées au sein de la première. Les séquences peptidiques des deux GFP sont indiquées ci-dessous :

>GFPuv <http://www.ncbi.nlm.nih.gov/gate1.inist.fr/protein/AAB06048.1>
mskgeelftg vvpilveldg dvnghkfsvs gegegdatyg kltlkfictt gklpvpwptl
vttfsygvqc fsrypdmkr hdffksampe gvyvqertisf kddgnyktra evkfegdtlv
nrielkgidf kedgnilgh leynynshnv yitadkqkng ikanfkirhn iedgsvqlad
hyqqntpig gpvllpdnhy lstqsalskd pnekrdhmv lefvtaagit hgmdelyk

>eGFP <http://www.ncbi.nlm.nih.gov/gate1.inist.fr/protein/ADQ73885.1>
mvskgeelft gvpilveld gdnghkfsv segegdaty gkltkfict gklpvpwpt
lvttltygvq cfsrypdmk qhdfksamp egyvqertif fkddgnyktr aevkfegdtl
vnrielkgid fkedgnilgh kleynynshn vymadkqkn gikvnfkirh niedgsvqla
dhyqqntpig dgpvllpdnh ylstsalsk dpnekrdhmv lefvtaagi tlgmdelyk

Souligné : partie N.GFP

Non souligné : partie C.GFP

En vert : le chromophore SYG (ou TYG)

En rouge : les différences compositionnelles entre les deux protéines

Ces deux GFP ont été utilisées comme rapporteur d'interaction.

a. 9.1 Autres exemples de « split-protéins »

Le système de criblage utilisant la split-GFP comme rapporteur d'interaction est un des exemples où le ré-assemblage des deux parties de la protéine initialement séparées, restitue la fonctionnalité de la protéine coupée, servant alors de témoin d'interaction des protéines fusionnées. La figure 17 ci-dessous montre d'autres exemples de protéines compatibles avec le système split-protéine [4].

La luciférase, par exemple, une fois reconstituée permet de localiser l'interaction entre les partenaires fusionnés grâce à sa couleur rouge. Son intérêt réside dans les avantages i) d'avoir peu de bruit de fond, du fait de sa liaison réversible contrairement à la GFP et ii) de ne pas avoir besoin de lumière exogène pour l'excitation et la révélation.

10. Stratégies

a. 10.1 Stratégie bi-cistronique

La stratégie « vecteur bi-cistronique » utilise un seul vecteur codant pour deux protéines de fusion. Un seul ARN messager est synthétisé et code pour deux protéines différentes sous le contrôle d'un seul promoteur (promoteur T7). Nous avons fait appel à la Société GenScript pour la construction de gènes synthétiques. Le vecteur dans lequel sont insérées les séquences d'intérêt est de type pUC. Le vecteur porte le gène qui code pour la GFPuv. La GFPuv est coupée en deux parties (NTER, aa 1 à 157) et (CTER aa 158 à 238). Chacune des parties est fusionnée à un partenaire.

Dans ce vecteur bi-cistronique, la partie N-terminale de la GFPuv (=N.GFP) est fusionnée à une cassette de clonage Gateway®. Cette cassette de clonage contient quatre séquences particulières: un gène léthal (ccdB), un gène de résistance au chloramphénicol (Chloramphenicol Acetyl Transferase) et deux sites de recombinaison (attR1 et attR2).

Quant à la partie C-terminale de la GFPuv (=C.GFP), elle est fusionnée à un des deux partenaires : MeV.XD ou SeV.XD. Nous avons alors deux plasmides nommés : pUC/SplitGFP-CcdB-MeV.XD et pUC/SplitGFP-CcdB-SeV.XD.

Le principe du clonage Gateway® repose sur une réaction de recombinaison, c'est-à-dire un échange de séquence d'ADN entre 2 séquences nucléotidiques. Les séquences à échanger sont balisées par des sites de recombinaison. Cet échange s'effectue précisément entre les sites de recombinaison et n'engendre aucune délétion ou insertion de nucléotide(s) (voir lien suivant : http://www.embl.de/pepcore/pepcore_services/cloning/cloning_methods/recombination/gateway/).

Cette stratégie s'est avérée techniquement inefficace pour obtenir les constructions contrôles et la banque de mutants. C'est la raison pour laquelle nous avons utilisé une deuxième stratégie : la stratégie deux vecteurs.

a. 10.2 Stratégie deux vecteurs

Cette stratégie s'appuie sur la publication "Detecting protein-protein interactions with GFP-fragment reassembly" [61]. La détection des interactions protéine-protéine est fondée sur le ré-assemblage des deux fragments de la eGFP codés chacun par un vecteur différent.

Dans ce cas, chaque partenaire d'un couple est inséré dans un vecteur différent. Chaque vecteur code pour une moitié de la eGFP. Le vecteur pET-N.GFP (AmpR) est sous le contrôle du promoteur T7, inductible avec l'iPTG, alors que le vecteur pMR-C.GFP (KanR) est sous le contrôle du promoteur pBAD, inductible avec l'arabinose. Nous avons demandé les vecteurs au Dr Lynne Regan, que nous avons par la suite transformés en cellule E.coli TAM1 pour réaliser une purification de plasmide (maxi-prep).

11. Vue globale du projet

Les protéines intrinsèquement désordonnées (PID) possèdent certaines propriétés relativement peu comprises, comme par exemple la capacité à lier de multiples partenaires, grâce à une interaction spécifique et réversible. Nous avons utilisé comme système modèle d'interaction et de reconnaissance de partenaire le couple NTAIL-PXD du virus de la rougeole afin d'explorer, sans a priori, les potentialités de reconnaissance moléculaire intrinsèques à la protéine désordonnée Ntail. Grâce à la mutagenèse aléatoire de la protéine Ntail et au système de criblage basé sur la fluorescence de la GFP reconstituée (sensible à l'affinité de l'interaction entre les partenaires mis en jeu), nous avons réalisé une évolution dirigée descriptive de la protéine Ntail. Compte tenu du fait que les substitutions amino-acidiques au sein de NTAIL vont modifier sa capacité à interagir avec son partenaire naturel PXD, les résultats obtenus vont permettre i) de préciser les résidus de NTAIL impliqués dans l'interaction, ii) d'apporter de nouveaux éclairages concernant les propriétés de reconnaissance vis-à-vis de son partenaire et iii) éventuellement de pouvoir dégager des concepts utiles pour les protéines désordonnées en général.

BIBLIOGRAPHIE

a. BIBLIOGRAPHIE

1. Uversky, V.N., C.J. Oldfield, and A.K. Dunker, *Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling*. J Mol Recognit, 2005. **18**(5): p. 343-84.
2. Kingston, R.L., et al., *Structural basis for the attachment of a paramyxoviral polymerase to its template*. Proc Natl Acad Sci U S A, 2004. **101**(22): p. 8301-6.
3. Tao, H. and V.W. Cornish, *Milestones in directed enzyme evolution*. Curr Opin Chem Biol, 2002. **6**(6): p. 858-64.
4. Shekhawat, S.S. and I. Ghosh, *Split-protein systems: beyond binary protein-protein interactions*. Curr Opin Chem Biol, 2011. **15**(6): p. 789-97.
5. Campos-Gonzalez, R. and J.R. Glenney, Jr., *Tyrosine phosphorylation of mitogen-activated protein kinase in cells with tyrosine kinase-negative epidermal growth factor receptors*. J Biol Chem, 1992. **267**(21): p. 14535-8.
6. Anfinsen, C.B. and E. Haber, *Studies on the reduction and re-formation of protein disulfide bonds*. J Biol Chem, 1961. **236**: p. 1361-3.
7. Kameshita, I. and H. Fujisawa, *Detection of protein kinase activities toward oligopeptides in sodium dodecyl sulfate-polyacrylamide gel*. Anal Biochem, 1996. **237**(2): p. 198-203.
8. Anfinsen, C.B., *Principles that govern the folding of protein chains*. Science, 1973. **181**(4096): p. 223-30.
9. Uversky, V.N., *Intrinsically disordered proteins from A to Z*. Int J Biochem Cell Biol, 2011. **43**(8): p. 1090-1103.
10. Babu, M.M., et al., *Intrinsically disordered proteins: regulation and disease*. Curr Opin Struct Biol, 2011. **21**(3): p. 432-40.
11. Dunker, A.K., et al., *Intrinsically disordered protein*. J Mol Graph Model, 2001. **19**(1): p. 26-59.
12. Uversky, V.N., *Intrinsically disordered proteins from A to Z*. Int J Biochem Cell Biol, 2011. **43**(8): p. 1090-103.
13. Lieutaud, P., B. Canard, and S. Longhi, *MeDor: a metaserver for predicting protein disorder*. BMC Genomics, 2008. **9 Suppl 2**: p. S25.
14. Callebaut, I., et al., *Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives*. Cell Mol Life Sci, 1997. **53**(8): p. 621-45.
15. Uversky, V.N. and A.K. Dunker, *Understanding protein non-folding*. Biochim Biophys Acta, 2010. **1804**(6): p. 1231-64.
16. Fink, A.L., *Compact intermediate states in protein folding*. Annu Rev Biophys Biomol Struct, 1995. **24**: p. 495-522.
17. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions*. Nat Rev Mol Cell Biol, 2005. **6**(3): p. 197-208.
18. Galea, C.A., et al., *Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits*. Biochemistry, 2008. **47**(29): p. 7598-609.
19. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm*. J Mol Biol, 1999. **293**(2): p. 321-31.
20. Uversky, V.N., *What does it mean to be natively unfolded?* Eur J Biochem, 2002. **269**(1): p. 2-12.
21. Stumpf, M.P., et al., *Estimating the size of the human interactome*. Proc Natl Acad Sci U S A, 2008. **105**(19): p. 6959-64.
22. Csizmok, V., et al., *Primary contact sites in intrinsically unstructured proteins: the case of calpastatin and microtubule-associated protein 2*. Biochemistry, 2005. **44**(10): p. 3955-64.
23. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder*. Nucleic Acids Res, 2003. **31**(13): p. 3701-8.
24. Fuxreiter, M., P. Tompa, and I. Simon, *Local structural disorder imparts plasticity on linear motifs*. Bioinformatics, 2007. **23**(8): p. 950-6.
25. Dunker, A.K. and Z. Obradovic, *The protein trinity--linking function and disorder*. Nat Biotechnol, 2001. **19**(9): p. 805-6.
26. Koshland, D.E., *Application of a Theory of Enzyme Specificity to Protein Synthesis*. Proc Natl Acad Sci U S A, 1958. **44**(2): p. 98-104.
27. Lee, H., et al., *Local structural elements in the mostly unstructured transcriptional activation domain of human p53*. J Biol Chem, 2000. **275**(38): p. 29426-32.
28. Tompa, P. and M. Fuxreiter, *Fuzzy complexes: polymorphism and structural disorder in*

- protein-protein interactions*. Trends Biochem Sci, 2008. **33**(1): p. 2-8.
29. Hazy, E. and P. Tompa, *Limitations of induced folding in molecular recognition by intrinsically disordered proteins*. Chemphyschem, 2009. **10**(9-10): p. 1415-9.
 30. Sigalov, A.B., *Uncoupled binding and folding of immune signaling-related intrinsically disordered proteins*. Prog Biophys Mol Biol, 2011. **106**(3): p. 525-36.
 31. Shoemaker, B.A., J.J. Portman, and P.G. Wolynes, *Speeding molecular recognition by using the folding funnel: the fly-casting mechanism*. Proc Natl Acad Sci U S A, 2000. **97**(16): p. 8868-73.
 32. Narayanan, R.L., et al., *Automatic assignment of the intrinsically disordered protein Tau with 441-residues*. J Am Chem Soc, 2010. **132**(34): p. 11906-7.
 33. Awile, O., et al., *Intrinsically disordered regions may lower the hydration free energy in proteins: a case study of nudix hydrolase in the bacterium Deinococcus radiodurans*. PLoS Comput Biol, 2010. **6**(7): p. e1000854.
 34. Edwards, Y.J., et al., *Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data*. Genome Biol, 2009. **10**(5): p. R50.
 35. Vavouri, T., et al., *Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity*. Cell, 2009. **138**(1): p. 198-208.
 36. Johansson, K., et al., *Crystal structure of the measles virus phosphoprotein domain responsible for the induced folding of the C-terminal domain of the nucleoprotein*. J Biol Chem, 2003. **278**(45): p. 44567-73.
 37. Bourhis, J.M., et al., *The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner*. Virus Res, 2004. **99**(2): p. 157-67.
 38. Gely, S., et al., *Solution structure of the C-terminal X domain of the measles virus phosphoprotein and interaction with the intrinsically disordered C-terminal domain of the nucleoprotein*. J Mol Recognit, 2010. **23**(5): p. 435-47.
 39. Lin, H. and V.W. Cornish, *Screening and selection methods for large-scale analysis of protein function*. Angew Chem Int Ed Engl, 2002. **41**(23): p. 4402-25.
 40. Wong, T.S., et al., *A statistical analysis of random mutagenesis methods used for directed protein evolution*. J Mol Biol, 2006. **355**(4): p. 858-71.
 41. You, L. and F.H. Arnold, *Directed evolution of subtilisin E in Bacillus subtilis to enhance total activity in aqueous dimethylformamide*. Protein Eng, 1996. **9**(1): p. 77-83.
 42. Kuchner, O. and F.H. Arnold, *Directed evolution of enzyme catalysts*. Trends Biotechnol, 1997. **15**(12): p. 523-30.
 43. Minshull, J., et al., *Engineered protein function by selective amino acid diversification*. Methods, 2004. **32**(4): p. 416-27.
 44. Moore, J.C., et al., *Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences*. J Mol Biol, 1997. **272**(3): p. 336-47.
 45. Brunet, A.P., et al., *The role of turns in the structure of an alpha-helical protein*. Nature, 1993. **364**(6435): p. 355-8.
 46. Ybe, J.A. and M.H. Hecht, *Sequence replacements in the central beta-turn of plastocyanin*. Protein Sci, 1996. **5**(5): p. 814-24.
 47. Fenniri, H., K.D. Janda, and R.A. Lerner, *Encoded reaction cassette for the highly sensitive detection of the making and breaking of chemical bonds*. Proc Natl Acad Sci U S A, 1995. **92**(6): p. 2278-82.
 48. Zhao, H. and F.H. Arnold, *Combinatorial protein design: strategies for screening protein libraries*. Curr Opin Struct Biol, 1997. **7**(4): p. 480-5.
 49. Black, M.E., et al., *Creation of drug-specific herpes simplex virus type 1 thymidine kinase mutants for gene therapy*. Proc Natl Acad Sci U S A, 1996. **93**(8): p. 3525-9.
 50. Schober, A., et al., *Accurate high-speed liquid handling of very small biological samples*. Biotechniques, 1993. **15**(2): p. 324-9.
 51. Baumgartner, J.W. and G.L. Hazelbauer, *Mutational analysis of a transmembrane segment in a bacterial chemoreceptor*. J Bacteriol, 1996. **178**(15): p. 4651-60.
 52. Zhou, H.X., R.H. Hoess, and W.F. DeGrado, *In vitro evolution of thermodynamically stable turns*. Nat Struct Biol, 1996. **3**(5): p. 446-51.
 53. Kast, P. and D. Hilvert, *3D structural information as a guide to protein engineering using genetic selection*. Curr Opin Struct Biol, 1997. **7**(4): p. 470-9.
 54. Cotten, S.W., et al., *mRNA display-based selections using synthetic peptide and natural protein libraries*. Methods Mol Biol, 2012. **805**: p. 287-97.
 55. Phillips, G.N., Jr., *Structure and dynamics of green fluorescent protein*. Curr Opin Struct Biol, 1997. **7**(6): p. 821-7.
 56. Miyawaki, A., T. Nagai, and H. Mizuno, *Mechanisms of protein fluorophore formation and engineering*. Curr Opin Chem Biol, 2003. **7**(5): p. 557-62.
 57. Remington, S.J., *Fluorescent proteins: maturation, photochemistry and photophysics*. Curr Opin Struct Biol, 2006. **16**(6): p. 714-21.
 58. Gross, L.A., et al., *The structure of the chromophore within DsRed, a red fluorescent*

- protein from coral*. Proc Natl Acad Sci U S A, 2000. **97**(22): p. 11990-5.
59. Yarbrough, D., et al., *Refined crystal structure of DsRed, a red fluorescent protein from coral, at 2.0-Å resolution*. Proc Natl Acad Sci U S A, 2001. **98**(2): p. 462-7.
 60. Wall, M.A., M. Socolich, and R. Ranganathan, *The structural basis for red fluorescence in the tetrameric GFP homolog DsRed*. Nat Struct Biol, 2000. **7**(12): p. 1133-8.
 61. Wilson, C.G., T.J. Magliery, and L. Regan, *Detecting protein-protein interactions with GFP-fragment reassembly*. Nat Methods, 2004. **1**(3): p. 255-62.
 62. Magliery, T.J., et al., *Detecting protein-protein interactions with a green fluorescent protein fragment reassembly trap: scope and mechanism*. J Am Chem Soc, 2005. **127**(1): p. 146-57.
 63. Ormo, M., et al., *Crystal structure of the Aequorea victoria green fluorescent protein*. Science, 1996. **273**(5280): p. 1392-5.
 64. Audic, S., et al., *SAmBA: an interactive software for optimizing the design of biological macromolecules crystallization experiments*. Proteins, 1997. **29**(2): p. 252-7.
 65. Gruet, A., S. Longhi, and C. Bignon, *One-step generation of error-prone PCR libraries using Gateway(R) technology*. Microb Cell Fact, 2012. **11**: p. 14.
 66. Simic, M., et al., *Driving forces of gyrase recognition by the addiction toxin CcdB*. J Biol Chem, 2009. **284**(30): p. 20002-10.
 67. Sarkar, M. and T.J. Magliery, *Re-engineering a split-GFP reassembly screen to examine RING-domain interactions between BARD1 and BRCA1 mutants observed in cancer patients*. Mol Biosyst, 2008. **4**(6): p. 599-605.
 68. Bourhis, J., et al., *The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner*. Virus Research, 2004. **99**: p. 157-67.
 69. Blocquel, D., et al., *Compaction and binding properties of the intrinsically disordered C-terminal domain of Henipavirus nucleoprotein as unveiled by deletion studies*. Mol Biosyst, 2012. **8**(1): p. 392-410.
 70. Bosley, A.D. and M. Ostermeier, *Mathematical expressions useful in the construction, description and evaluation of protein libraries*. Biomol Eng, 2005. **22**(1-3): p. 57-61.