

Thèse de Doctorat

affilogic

Simon HUET

*Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le label de L'Université Nantes Angers Le Mans*

École doctorale : *Biologie-Santé (ED 502)*

Discipline : *Biomolécules, pharmacologie, thérapeutique*

Unités de recherche : **Unité Fonctionnalité et Ingénierie des Protéines**
UMR CNRS 6286
UFR des Sciences et Techniques
44322 Nantes

Affilogic SAS
2 rue de la Houssinière
44300 Nantes

Soutenue le 30 Octobre 2015

Thèse CIFRE N° : 2012/1337

Reconnaissance moléculaire : De l'*in silico* à l'*in vitro* et vice versa

JURY

Rapporteurs : **Thomas SIMONSON**, Directeur de Recherche et Professeur Associé, Ecole Polytechnique
Vincent DIVE, Directeur de Recherche, Commissariat à l'Energie Atomique

Examineurs : **Philippe MINARD**, Professeur des Universités, Université Paris Sud
Leonardo SCAPOZZA, Professeur des Universités, Université de Genève
Yves-Henri SANEJOUAND, Directeur de Recherche, Université de Nantes
Mathieu CINIER, Directeur Scientifique, Affilogic

Directeur de Thèse : **Yves-Henri SANEJOUAND**, Directeur de Recherche, Université de Nantes

Co-encadrant de Thèse : **Mathieu CINIER**, Directeur Scientifique, Affilogic

Remerciements

Mes remerciements s'adressent tout d'abord aux membres du jury, qui ont accepté d'évaluer les travaux réalisés depuis Janvier 2013. J'exprime ma reconnaissance envers Thomas Simonson et Vincent Dive pour avoir examiné le présent manuscrit. Merci également à Philippe Minard et Leonardo Scapozza pour le suivi qu'ils ont assuré au cours de la thèse, et pour la pertinence dont ils ont fait preuve lors de nos rencontres. Enfin, je remercie Yves-Henri Sanejouand et Mathieu Cinier pour l'évaluation finale des travaux qu'ils ont tous deux dirigés.

Sans pouvoir dresser de liste exhaustive, je tiens à remercier les membres des laboratoires qui m'ont accueilli. En premier lieu, un énorme merci à toute l'équipe d'Affilogic dans laquelle je me sens à ma place et grâce à qui travailler est un véritable plaisir. “!Muchas gracias!” à Olivier Kitten pour l'opportunité qu'il m'a offerte, sa confiance, son soutien pré-, per- et post-thèse, et pour rendre l'aventure Affilogic possible. Dans le désordre, “Thk” à Mathieu Cinier pour son encadrement et ses conseils, mais aussi les fléchettes dans les côtes, les séances McDo, les pizzas déstructurées ou les roulades grâce auxquelles nos liens dépassent désormais le cadre strictement professionnel. J'adresse aussi un “You rock” à Nadège Prel pour assurer autant à l'escalade, au laboratoire ou en dehors, mais aussi pour la poussée d'Archimède (dans mon répertoire musical du moins). Enfin, un “Ci-mer” à l'ensemble de l'équipe, complétée par Anaëlle, Anne, Ariane, Astrid, Chloé, Harmony, Jessy, Justine, Léo, Nathalie et Stéphane par ordre alphabétique, chacun à leur tour moteurs d'Affilogic et équipier(ère)s plaisant(e)s. En plus de l'environnement de travail épanouissant, merci pour les séances de footing, de jeux de société, de blagues à 0,50 cents (trois francs six sous, inflation non comprise), de films de tornades de requins, et d'innombrables bugs informatiques à résoudre.

Ma reconnaissance se dirige vers l'équipe de l'UFIP (et de l'U3B!) qui m'a donné l'envie et l'opportunité de réaliser une thèse à l'interface entre biotechnologies et bioinformatique. Tout d'abord, merci à Yves-Henri Sanejouand d'avoir dirigé mes recherches durant la thèse, de m'avoir accordé sa confiance et ses recommandations, me permettant de découvrir un domaine de la modélisation moléculaire que je n'avais jusqu'alors jamais exploré. Je remercie également

Charles Tellier d'avoir suscité chez moi un attrait pour les analyses d'interactions biomoléculaires et les biotechnologies en règle générale, en plus de son soutien et ses conseils pendant la thèse. Enfin, merci aux membres de l'UFIP qui ont participé à ma formation et m'ont offert leur soutien moral, notamment lors des divers rassemblements intra- et extra-laboratoire.

Je suis reconnaissant envers Leonardo Scapozza et les membres de son équipe à l'Université de Genève, qui m'ont accueilli lors d'un criblage de conditions de cristallogénèse. En particulier, merci à Andreja Vujicic Zagar et Magali Zeisser Labouèbe pour m'avoir fait découvrir le laboratoire et agréablement accompagné lors de mon séjour.

Au delà de relations professionnelles, j'ai été pris d'une étrange sensation au cours de ces trois années auprès de ceux que je considère comme des amis. Si ce n'est pas une indigestion de sushis, alors ce doit être de la gratitude. En plus des personnes citées précédemment qui se reconnaîtront, un gigantesque merci aux exilés dont l'absence s'est faite sentir: Marine Goux et Denis Velic. Nul doute que vous auriez été du même soutien sans faille sans avoir à gérer votre propre thèse. Heureux que vous soyez de retour! Merci également à Maxime Julien et David Tezé, aspirés par la chimie, d'avoir participé à mon envie de démarrer un projet de thèse ainsi qu'à l'augmentation régulière de mon taux de $\text{CH}_3\text{-CH}_2\text{-OH}$ depuis leur rencontre. Merci à ceux que je taquine suffisamment pour qu'ils puissent se considérer comme partie intégrante de mes proches. Enfin, j'exprime ma gratitude envers ceux en qui je place ma confiance, et à qui j'offre mes pensées sans détour et sans délicatesse en guise de "Merci". A Georges Abitbol.

Plus intimement, je suis extrêmement reconnaissant envers ma famille (en particulier mes parents Gérald et France, mon frère Alexandre, ma sœur Romane, et mes grand-parents) mais aussi la famille avec laquelle je ne partage pas de lien de sang. Je ne les remercierai jamais assez pour leur présence, leur soutien moral et tout ce qu'ils ont mis en œuvre pour que je puisse finalement rédiger ces lignes.

Pour finir, merci à tous les problèmes à affronter, comme le conseillait Jim Lovell, car quelqu'un avec moins de compétences pourrait avoir mon travail s'il étaient moins difficiles.

Epigraphe

« 42 »

Super-calculateur *Pensées Profondes* en réponse à la Grande Question sur la Vie, l'Univers et le Reste

Après 7,5 millions d'années de calculs — *H2G2, Le Guide du Voyageur Galactique*

Financements

Les travaux présentés dans le présent mémoire ont été réalisés dans le cadre d'une thèse en convention industrielle de formation par la recherche (CIFRE, convention N°2012/1337) cofinancée par le Fonds européen de développement régional (FEDER, convention 2013/FEDER/254 référence 39245) dans l'Unité Fonctionnalité et Ingénierie des Protéines (UFIP), laboratoire UMR CNRS 6286, et au sein d'Affilogic, société nantaise créée en Février 2010.



Cette thèse est cofinancée par l'Union Européenne. L'Europe s'engage en Pays de la Loire avec le Fonds européen de développement régional.

Table des matières

Remerciements	3
Epigraphe	5
Financements	7
Table des matières	9
Liste des abréviations et acronymes	15
Avant-propos	19
Chapitre I: Environnements scientifiques et méthodologiques	23
<i>1.1. Les protéines d'affinité à usage thérapeutique</i>	23
1.1.1. Un concept dérivé des anticorps	23
1.1.2. Limitations des anticorps et développement d'alternatives	24
<i>1.2. Ingénierie des protéines et apports des outils de modélisation</i>	29
1.2.1. Optimisations au service du développement de médicaments	29
1.2.2. Stabilité et solubilité.....	31
1.2.3. Demi-vie	32
1.2.4. Immunogénicité	32
1.2.5. Repliement et fonction	34
1.2.6. Applications aux Nanofitines, protéines d'affinité développées à Affilogic.....	34
<i>1.3. Transfert de fonction et stratégie d'humanisation</i>	37
1.3.1. Pourquoi humaniser les protéines à usage thérapeutique ?.....	37
1.3.2. Succès de greffes protéiques	38
1.3.2.1. Sites de coordination d'ions métalliques.....	39
1.3.2.2. Activités enzymatiques	40
1.3.2.3. Modèles de protéines membranaires.....	41
1.3.3. Greffes d'interfaces protéine-protéine	42
1.3.3.1. Transferts de boucles: humanisation classique d'anticorps.....	42
1.3.3.2. Transferts de boucles: miniaturisation de charpentes	43
1.3.3.3. Transferts depuis des interacteurs naturels	45
1.3.3.4. Transferts de feuillets beta.....	48
<i>1.4. Approches computationnelles de design rationnel</i>	49
1.4.1. Protéines de liaison et reconnaissance moléculaire spécifique	49
1.4.2. État de l'art en design protéique <i>in silico</i>	54
1.4.2.1. Généralités	54
1.4.2.2. Méthodes disponibles et exemples	55
1.4.2.3. L'expérience CAPRI	58

I.4.3. La suite de modélisation Rosetta.....	61
I.4.3.1. Forces et améliorations apportées aux modèles prédits.....	61
I.4.3.2. Résolution de structure.....	62
I.4.3.3. Ancrage moléculaire	65
I.4.3.4. Re-design.....	67
I.4.3.5. Design <i>de novo</i>	70
Chapitre II: Humanisation de Nanofitines par greffe de domaine	79
II.1. Stratégie d'humanisation des Nanofitines	79
II.1.1. Des domaines de liaison exposés sur le feuillet beta de Sac7d.....	79
II.1.2. Enseignements du transfert d'un site de Nanofitine entre feuillet et boucles	80
II.1.3. Transfert du feuillet anti-GFP de Nanofitines.....	81
II.2. Découverte de Nanofitines anti-GFP interagissant via leur feuillet	83
II.2.1. Choix de la banque de variants de Nanofitines	83
II.2.2. Sélection des Nanofitines anti-GFP par <i>ribosome display</i>	86
II.2.3. Choix de 2 domaines de Nanofitines anti-GFP à transférer	87
II.2.3.1. Critères de choix de Nanofitines anti-GFP	87
II.2.3.2. Expression et purification des Nanofitines.....	88
II.2.3.3. Détermination des constantes d'affinité.....	89
II.2.3.4. Identification d'épitopes distincts sur la GFP.....	90
II.2.3.5. Sites de fixation à la GFP choisis pour le transfert de charpente	92
II.3. Charpentes protéiques humaines hôtes de feuillet anti-GFP	92
II.3.1. Recherche structurale de feuillets hôtes.....	92
II.3.2. Comparaison avec les structures en <i>OB-fold</i>	95
II.3.3. Sélection des charpentes hôtes.....	96
II.4. Expression sous forme soluble des charpentes humaines.....	98
II.4.1. Obtention et clonage des séquences codantes	98
II.4.2. Expression en bactérie et impact du transfert sur la solubilité	100
II.4.2.1. Cultures standards	100
II.4.2.2. Exploration de conditions augmentant la solubilité	101
II.4.3. Impact du transfert sur les structures 1C8C et 4F7H.....	104
II.5. Conclusions et observations sur la méthodologie de transfert de feuillets.....	106
II.5.1. Limitations mises en avant expérimentalement.....	106
II.5.2. Prédictions complémentaires implémentées <i>a posteriori</i>	106
II.5.2.1. Hydrophobie	107
II.5.2.2. Stabilité.....	109
II.5.2.3. Diversité des profils prédits	110
II.5.2.4. Cartographie de l'interaction Nanofitine C8 - GFP	111

II.5.3. Conclusions et perspectives d'améliorations méthodologiques	115
Chapitre III: Design de novo de Nanofitine spécifique	121
III.1. Stratégie de design de novo de Nanofitines anti-GFP	121
III.1.1. Complémentarité des approches <i>in silico</i> et <i>in vitro</i>	121
III.1.2. Buts et stratégie de design de Nanofitines	122
III.2. Obtention et préparation des poses initiales.....	124
III.3. Ancrage moléculaire pour la recherche de complémentarité spatiale.....	127
III.3.1. Description de l'algorithme	127
III.3.2. Impact des séquences des Nanofitines.....	128
III.3.3. Découverte de régions et d'orientations géométriques favorables au ciblage	130
III.3.4. Orientation des Nanofitines pour la formation d'interfaces	134
III.3.5. Sélection des complexes pour le design protéique	139
III.4. Design protéique: Optimisation de l'interface par ingénierie de la surface de la Nanofitine	141
III.4.1. Description de l'algorithme	141
III.4.2. Convergence vers des interfaces Nanofitine:GFP étendues.....	142
III.4.3. Obtention de 10 potentielles Nanofitines anti-GFP.....	145
III.5. Confirmations expérimentales d'une fixation spécifique à la GFP	147
III.5.1. Criblage des Nanofitines anti-GFP prédites	147
III.5.2. Validation de l'épitope ciblé par NF5	150
III.6. Optimisation de l'affinité de NF5 par modélisation.....	155
III.6.1. Stratégie d'exploration au voisinage de la pose NF5:GFP	155
III.6.2. Convergence vers la famille de NF1 à NF5	156
III.7. Conclusions et perspectives du design de novo de Nanofitines	158
III.7.1. Objectifs atteints	158
III.7.2. Extension des paramètres modélisés	160
III.7.3. Approches alternatives	162
Chapitre IV: Conclusions générales et perspectives.....	167
IV.1. Ingénierie rationnelle des Nanofitines	167
IV.2. Etat des avancées réalisées	168
IV.2.1. Transfert de site de liaison pour l'humanisation de Nanofitines.....	168
IV.2.2. Design rationnel assisté par modélisation.....	169
IV.3. Défis et perspectives	170
Chapitre V: Matériels et Méthodes	177
V.1. Modélisations moléculaires.....	177
V.1.1. Recherche de feuillettes compatibles dans la PDB.....	177

V.1.2. Analyses d'hydrophobie	177
V.1.3. Calculs de stabilité.....	178
V.1.3.1. Description de la méthode	178
V.1.3.2. Exécution des simulations.....	178
V.1.4. Design <i>de novo</i>	179
V.1.4.1. Préparation des poses	179
V.1.4.2. Ancrage moléculaire.....	180
V.1.4.3. Regroupement des poses.....	182
V.1.4.4. Design de la surface d'interaction des Nanofitines	184
V.1.4.5. Aires de contact et cavités à l'interface	187
V.1.4.6. Design complémentaire (après le tour initial).....	188
V.2. Validations expérimentales.....	190
V.2.1. Liste des réactifs.....	190
V.2.1.1. Liste des oligonucléotides utilisés	190
V.2.1.2. Autres réactifs	194
V.2.2. Sélection <i>in vitro</i> de Nanofitines anti-GFP	196
V.2.2.1. Biotinylation des antigènes	196
V.2.2.2. Tours de sélection par <i>ribosome display</i> et isolement des clones	196
V.2.2.3. Criblage des Nanofitines anti-GFP par ELISA.....	197
V.2.2.4. Criblage complémentaire par interférométrie de couche biologique	197
V.2.3. Constructions de plasmides.....	198
V.2.3.1. Isolement, amplification et purification de plasmides.....	198
V.2.3.2. Sous-clonage des gènes codant pour les charpentes sauvages et greffées.....	199
V.2.3.3. Construction des plasmides pAFG12, pAFG19, pAFG20, pAFG21	201
V.2.3.4. Construction de Nanofitines ponctuelles (NF1 à NF10, variants alanine de C8).....	202
V.2.3.5. Changement de résistance du plasmide pAFG05-GFP	204
V.2.3.6. Construction de tétramères de Nanofitine.....	205
V.2.3.7. Construction de Nanofitine sans étiquette de fusion.....	207
V.2.3.8. Construction dédiées à la cristallogénèse	207
V.2.3.9. Mutagenèse en alanine des variants de la GFP	208
V.2.4. Expression et purification des protéines recombinantes.....	209
V.2.4.1. Cultures bactériennes.....	209
V.2.4.2. Cultures bactériennes alternatives	209
V.2.4.3. Purification	210
V.2.4.4. Dénaturation-renaturation et purification des protéines insolubles	212
V.2.4.5. Traduction <i>in vitro</i>	212
V.2.5. Caractérisations des protéines exprimées	214

V.2.5.1. SDS-PAGE	214
V.2.5.2. ELISA	214
V.2.5.3. Interférométrie de couche biologique	215
V.2.5.4. Mesures de fluorescence.....	217
V.2.5.5. Criblage des conditions de cristallogénèse	218
Bibliographie	221
Annexes.....	245
Annexe 1. Scripts.....	245
Annexe 1.1. queue.sh	245
Annexe 1.2. scores.sh	245
Annexe 1.3. preparation.sh	246
Annexe 1.3.1. Script de préparation des fichiers resfile	246
Annexe 1.3.2. Modèle pour 1AZP	246
Annexe 1.3.3. Modèle pour 1C8C	246
Annexe 1.3.4. Modèle pour 1LPJ.....	247
Annexe 1.3.5. Modèle pour 1QNT.....	247
Annexe 1.3.6. Modèle pour 4F7H	247
Annexe 1.4. interactions-clusters.py.....	247
Annexe 1.5. pymol-CB-rms.py.....	251
Annexe 1.6. interactions-resn-to-clusters.py.....	252
Annexe 1.7. weblogo-loop.sh	255
Annexe 1.8. pdbsurf-auto.sh (v1.3).....	256
Annexe 1.9. batch.sh	261
Annexe 1.10. pdbsurf-auto.sh (v1.5)	275
Annexe 1.11. pdb2fasta.sh	280
Annexe 1.12. pdbsurf-batch.sh	282
Annexe 1.13. crontab.sh	285
Annexe 2. Productions scientifiques	286
Annexe 2.1. Publication	286
Annexe 2.2. Communications orales	287
Annexe 2.3. Posters	289

Liste des abbréviations et acronymes

- 10Fn3: Dixième domaine de la fibronectine de type III
- 1AZP: Identifiant PDB de la structure de référence pour Sac7d
- 1C8C: Identifiant PDB de la structure de référence pour Sso7d
- 1LPJ: Identifiant PDB de la structure de référence pour une lipocaline, hôte du transfert de site
- 1QNT: Identifiant PDB de la structure de référence pour une protéine de réparation de l'ADN, hôte du transfert de site
- 3D: Tridimensionnel(le)
- 3LVA: Identifiant PDB de la structure de référence pour la GFP
- 4F7H: Identifiant PDB de la structure de référence pour une fermitine, hôte du transfert de site
- Å: Ångström
- A3: Nanofitine anti-GFP A3
- AA: Acides aminés
- ADN: Acide désoxyribonucléique
- app: Polypeptide pancréatique aviaire
- ARN : Acide ribonucléique
- ARNm: ARN messenger
- B4: Nanofitine anti-GFP B4
- BSA: Albumine de sérum bovin, ou *Bovine Serum Albumin*
- C2: Nanofitine anti-GFP C2
- C5: Nanofitine anti-GFP C5
- C6: Nanofitine anti-GFP C6
- C8-7-GFP, C8-10-GFP, C8-20-GFP: C8 fusionnée à la GFP via un bras espaceur de 7, 10 ou 20 résidus
- C8: Nanofitine anti-GFP C8
- CA→C: Vecteur entre le carbone alpha et le carbone du groupe carboxyle d'un résidu acide aminé
- CAPRI: *Critical Assessment of Predicted Interactions*
- CASP: *Critical Assessment of Techniques for Protein Structure Prediction*
- CDR-H3: Troisième boucle de CDR de chaîne lourde
- CDR: Région déterminant la complémentarité, ou *Complementary Determining Region*
- CIFRE: Convention industrielle de formation par la recherche
- cos: Cosinus
- CTLA-4: Molécule de co-stimulation négative, ou *Cytotoxic T-Lymphocyte-Associated protein 4*
- D8: Nanofitine anti-GFP D8
- DsbC: Protéine chaperonne, ou *Disulfide bond C isomerase*
- E. coli: Escherichia coli*
- E2: Nanofitine anti-GFP E2
- EC₅₀: Concentration efficace médiane

EGFR: Récepteur au facteur de croissance épidermique, ou *Epidermal Growth Factor Receptor*

EL: Fraction purifiée et éluée par chromatographie

ELISA: Méthode immuno-enzymatique ELISA, ou *Enzyme-Linked Immunosorbent Assay*

EPO: Erythropoïétine

Fab: Fragment Fab d'anticorps, ou *Fragment Antigen-Binding region*

Fc: Fragment cristallisable d'anticorps, ou *Fragment crystallizable region*

FEDER: Fonds européen de développement régional

f_{nat} : Fraction des contacts natifs inter-résidus

g : Unité d'accélération

GFP: Protéine fluorecente verte, ou *Green Fluorescent Protein*

GFP Δ linker: GFP délétée du bras espaceur superflux

H4: Nanofitine anti-lysozyme H4

HER2: Récepteur d'hormones de croissance épidermique humain HER2, ou *Human Epidermal growth factor Receptor-2*

hGH: Hormone de croissance humaine , ou *human Growth Hormone*

hGHR: Récepteur à l'hormone de croissance humaine , ou *human Growth Hormone Receptor*

HRP: Peroxydase de raifort, ou *HorseRadish Peroxidase*

I_rms: RMSD des carbones alpha de l'interface

I_sc: Score d'interaction

IgG: Immunoglobuline de type G

IL-12: Interleukine 12

IL-17: Interleukine 17

IL-23: Interleukine 23

IMAC: Chromatographie d'affinité sur métal immobilisé, ou *Immobilized Metal ion Affinity Chromatography*

K_D : Constante de dissociation à l'équilibre

kDa: Kilodalton

K_i : Constante d'inhibition

k_{off} : Constante cinétique de dissociation

k_{on} : Constante cinétique d'association

L_rms: RMSD de la position des carbones alpha du partenaire le plus petit après superposition de la protéine de plus grande taille

l: Litre

lon: Protéase de *E. coli*

mAbs: anticorps monoclonaux, ou *monoclonal Antibodies*

MBP: Protéine de liaison au maltose, ou *Maltose Binding Protein*

mg: Milligramme

MHC-II: Complexes majeurs d'histocompatibilité de classe II

ml: Millilitre

mM: Millimolaire

N→C: Vecteur entre l'azote du groupe amine et le carbone du groupe carboxyle d'un résidu acide aminé

N→CA: Vecteur entre l'azote du groupe amine et le carbone alpha d'un résidu acide aminé

NF1 à NF10: Nanofitines générées *in silico*

NF5-7-GFP, NF5-10-GFP, NF5-20-GFP: NF5 fusionnée à la GFP via un bras espaceur de 7, 10 ou 20 résidus

NF5x4, NF7x4, NF9x4: Tétramères linéaires de Nanofitines générées *in silico*

ng: Nanogramme

nm: Nanomètre

nM: Nanomolaire

NR: Fraction non retenue par chromatographie

NusA: Facteur de transcription NusA, ou *N-utilising substance A*

OB-fold: Repliement en motif de fixation aux oligonucléotides/oligosaccharides, ou (*Oligonucleotide/Oligosaccharide Binding*)-fold

ompT: Protéase membranaire de *E. coli*

pAFG01: Plasmide d'expression avec fusion N-terminale d'une étiquette RGS-His₆

pAFG05-GFP: Plasmide d'expression de la GFP avec fusion N-terminale d'une étiquette Strep-TagII

pAFG16: Plasmide d'expression avec fusion C-terminale d'une étiquette RGS-His₆

pAFG19: Plasmide d'expression avec fusion N-terminale de TrxA et d'un site de clivage TEV, et fusion C-terminale d'une étiquette RGS-His₆

pAFG20: Plasmide d'expression avec fusion N-terminale de NusA et d'un site de clivage TEV, et fusion C-terminale d'une étiquette RGS-His₆

pAFG21: Plasmide d'expression avec fusion N-terminale de MBP et d'un site de clivage TEV, et fusion C-terminale d'une étiquette RGS-His₆

PCR: Réaction en chaîne par polymérase, ou *Polymerase Chain Reaction*

PDB: Banque de données sur les protéines, ou *Protein Data Bank*

PEG: Polyéthylène glycol

pEX-A: Plasmide de clonage de gènes synthétiques

pKa: Constante logarithmique d'acidité

pM: Picomolaire

pmol: Picomole

R: Résolution

R²: Coefficient de détermination

RCPG: Récepteur couplé aux protéines G

rfu: Unités relative de fluorescence, ou *relative fluorescence units*

RMS: Racine de la moyenne des carrés, ou *Root Mean Square*

RMSD: Déviation de la racine de la moyenne des carrés, ou *Root Mean Square Deviation*

rpm: Rotations par minute

RT-PCR: Réaction en chaîne par polymérase après transcription inverse, ou *Reverse Transcription Polymerase Chain Reaction*

Sac7d: Protéine de liaison à l'ADN de 7 kDa, découverte chez *Sulfolobus acidocaldarius*

scFv: Fragment variable à chaîne unique, ou *single-chain variable fragment*

SDS-PAGE: Electrophorèse sur gel de polyacrylamide en présence de dodécylsulfate de sodium, ou *Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis*

SEC: Chromatographie d'exclusion stérique, ou *Size-exclusion chromatography*

Sso7d: Protéine de liaison à l'ADN de 7 kDa, découverte chez *Sulfolobus solfataricus*

TEV: Protéase du virus de la gravure du tabac, ou *Tobacco Etch Virus protease*

Tm: Température de fusion, ou *melting temperature*

TNF α : Facteur de nécrose tumorale alpha, ou *Tumor Necrosis Factor alpha*

TPP: profil de produit cible, ou *Target Product Profile*

TrxA: Thioredoxine A

UFIP: Unité Fonctionnalité et Ingénierie des Protéines, laboratoire UMR CNRS 6286

VEGF: Facteur de croissance de l'endothélium vasculaire, ou *Vascular Endothelial Growth Factor*

VHH: chaîne lourde variable d'anticorps à chaîne lourde de camélidés, ou *heavy chain variable domain in camelids*

VIH: Virus de l'immunodéficience humaine

WT: sauvage, ou *wild-type*

λ_{\max} : longueur d'onde maximale

μ g: Microgramme

μ l: Microlitre

μ M: Micromolaire

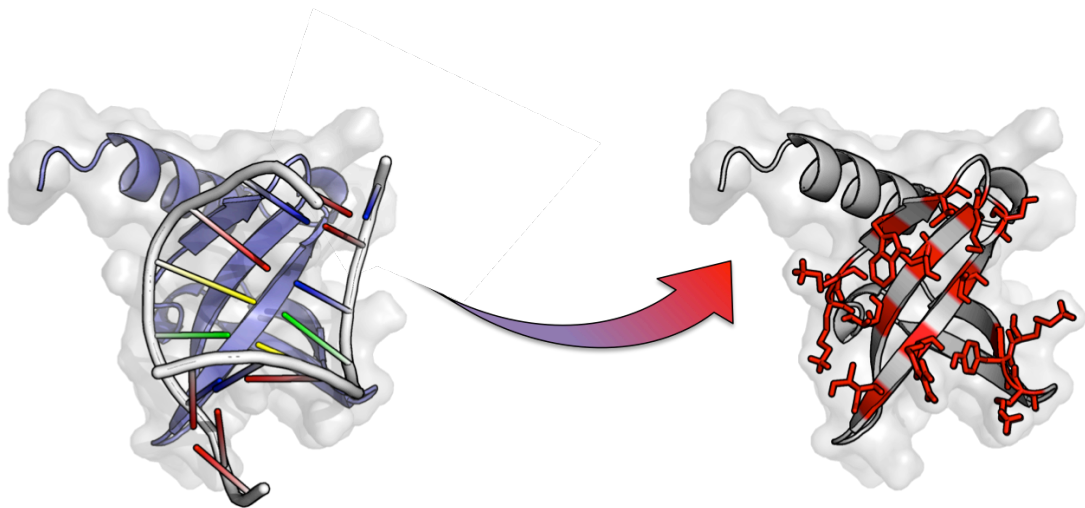
Avant-propos

“Reconnaissance moléculaire, de l’*in silico* à l’*in vitro* et *vice versa*”. Derrière ce titre, se cachent l’objet principal des études décrites dans ce manuscrit, ainsi que l’usage conjoint de techniques pluridisciplinaires d’ingénierie des protéines à travers des aspects de biologie moléculaire, de biochimie, et de bioinformatique.

Une dimension du travail de thèse que nous aborderons ici appartient au monde de l’*in silico*. Référence aux circuits imprimés à base de silicium, ce néologisme d’inspiration latine regroupe les outils informatiques dont les modèles mathématiques sont exploités en recherche, et a été cité dans la littérature scientifique pour la première fois en 1991 (*Hansen et al., 1991*). Cette facette est à distinguer – mais aussi à compléter – par la dimension *in vitro* de ce manuscrit, littéralement “dans le verre” en latin, désignant les expérimentations réalisées en tubes à essai de laboratoire. Ces dimensions *in silico* et *in vitro* ont été explorées en quête de la modélisation et du contrôle de la reconnaissance moléculaire, représentant la formation spécifique de complexes entre molécules par l’intermédiaire d’interactions physico-chimiques. Dans notre cas, ce phénomène a été étudié entre protéines, avec comme partenaire central une nouvelle classe de protéines d’affinité destinée à développer les médicaments de demain: les Nanofitines.

Poussés par la volonté de guider les approches du verre par celles de la silice (et *vice versa*) dans une collaboration mutuelle au service du perfectionnement méthodologique et de l’optimisation des produits de nos recherches, nous décrivons successivement dans ce manuscrit deux approches expérimentales et prédictives: une stratégie d’humanisation de Nanofitines, puis leur design *de novo*.

Chapitre I: Environnements scientifiques et méthodologiques



Chapitre I: Environnements scientifiques et méthodologiques

1.1. Les protéines d'affinité à usage thérapeutique

1.1.1. Un concept dérivé des anticorps

Depuis l'approbation de l'insuline en 1982 (Goeddel et al., 1979), les protéines recombinantes ont pris une part grandissante dans l'arsenal des composés actifs à usage thérapeutique. En 2010, les protéines thérapeutiques, dont 48% étaient des anticorps monoclonaux (mAbs), ont représenté un marché de plus de 100 milliards de dollars rien que pour l'Union Européenne et les Etats-Unis (Dimitrov, 2012). Parmi les protéines thérapeutiques les plus vendues cette année-là, on dénombre 6 mAbs et une protéine de fusion avec un fragment Fc. En juillet 2011, les mAbs et les protéines de fusion avec un fragment Fc approuvés comme médicaments ont été portés respectivement au nombre de 29 et 6 (Dimitrov, 2012), représentant 30% des autorisations de mise sur le marché pour des mAbs atteignant 5% des médicaments vendus. Tous ces chiffres mettent en évidence l'impact économique croissant (Moran, 2011) et la position dominante des anticorps dans les domaines du diagnostic et du traitement de pathologies (cancer et inflammation essentiellement, **Figure 1**).

Afin de produire l'effet thérapeutique désiré, différents mécanismes biologiques sont ciblés par les anticorps thérapeutiques actuellement commercialisés, dont les interactions de type récepteur-ligand et leurs cascades de signalisation associées. En effet, le dérèglement de ces processus biologiques est souvent la cause de pathologies, les plaçant donc comme cible d'intérêt pour le développement de nouveaux médicaments. Par exemple, les mAbs Cetuximab et Panitumumab inhibent l'activation du récepteur au facteur de croissance épidermique (EGFR) par son ligand. La fixation de l'anticorps sur EGFR bloque l'accès de son agoniste (Sunada et al., 1986) et empêche sa dimérisation (nécessaire à la transduction du signal; S. Li et al., 2005). Cette propriété est à l'origine de leur indication pour le cancer colorectal (Weiner et al., 2010). De même, le Bevacizumab bloque la fixation du facteur de croissance de l'endothélium vasculaire à ses récepteurs et a été approuvé dans le traitement des cancers colorectaux, pulmonaires et de la poitrine (Ellis et Hicklin, 2008). A la frontière des anticorps, le Belatacept est une protéine de

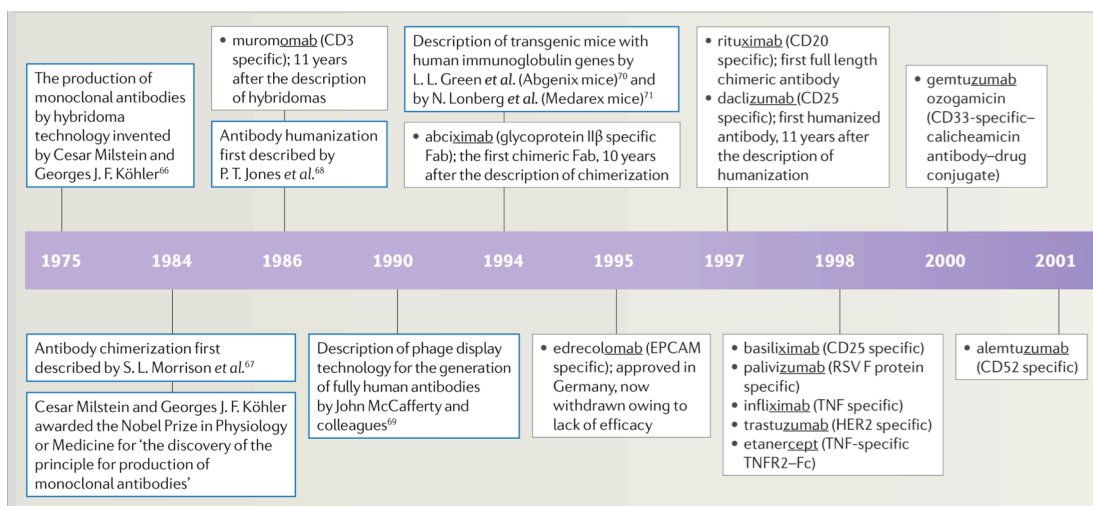


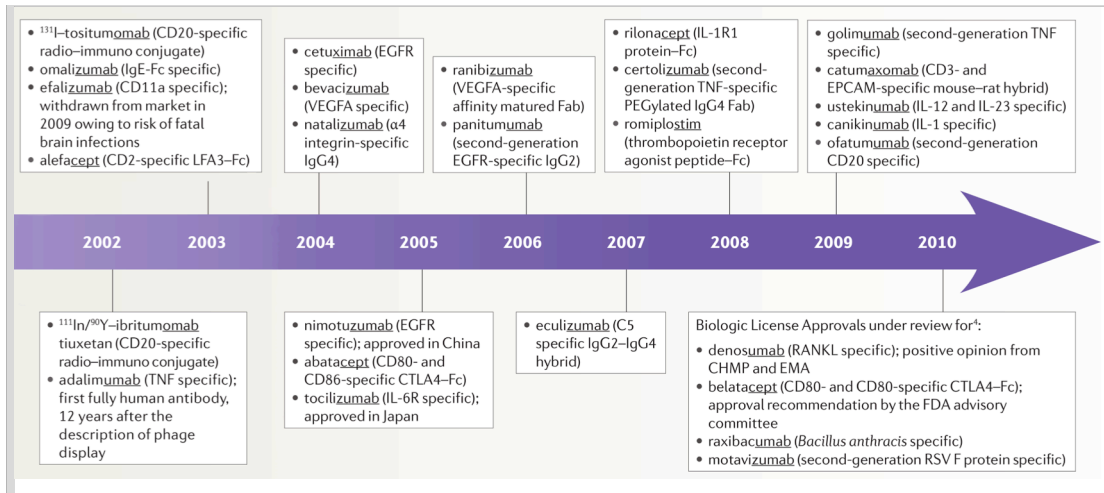
Figure 1 : Principales évolutions des techniques et « Blockbusters » dérivés d'IgG approuvés au cours des 30 dernières années. Les suffixes des molécules indiquent le format de l'anticorps. C5, complément component C5; CHPM, Committee for Medicinal Products for Human Use; CTLA4, cytotoxic T lymphocyte antigen 4; EGFR, epidermal growth factor receptor; EPCAM, epithelial cell adhesion molecule; HER2, human epidermal growth

fusion entre un Fc modifié et la molécule de costimulation CTLA-4 développée pour inhiber l'interaction entre les molécules B7 de cellules présentatrices d'antigènes et le récepteur CD28 des lymphocytes T. La saturation des récepteurs de co-stimulation B7-1 et B7-2 inhibe la voie induite par CD28, qui est critique pour l'activation des cellules T (Linsley et Ledbetter, 1993). Cela justifie l'utilisation du Belatacept comme immunosuppresseur lors de transplantations rénales (Wekerle et Grinyó, 2012).

I.1.2. Limitations des anticorps et développement d'alternatives

Les anticorps demeurent pour beaucoup le modèle de choix parmi les protéines d'affinité, notamment pour le développement d'agents neutralisants à vocation thérapeutique. Cette position est confortée par la grande expérience quant à leur utilisation et par le succès de leurs versions recombinantes ou humanisées dans des applications cliniques.

Toutefois, les anticorps souffrent de nombreuses contraintes inhérentes à leur structure (Binz et al., 2005). Composés de plusieurs domaines structurés par des ponts disulfures et soumis à des modifications post-traductionnelles (glycosylations), leur production est jugée relativement difficile et onéreuse (Steinmeyer et McCormick, 2008). Les anticorps sont également sensibles à la dégradation, à l'agrégation, aux modifications (oxydation ou désamination par exemple) et à la dénaturation (Ruigrok et al., 2011). Bien que des versions plus petites comme les fragments scFv ou Fab aient été produites avec succès dans des systèmes microbiens, leur stabilité repose



factor receptor 2; IL, interleukin; IL-IRAP, IL-IR accessory protein; LFA3, lymphocyte function-associated antigen 3; PEG, polyethylene glycol; R, receptor; RANKL, receptor activator of nuclear factor- κ B ligand; RSV, respiratory syncytial virus; TNF, tumour necrosis factor; TNFR2, TNF receptor 2; VEGFA, vascular endothelial growth factor A. Adapté d'après Beck et al., 2010.

toujours sur la création de ponts disulfures intradomains qui ne peuvent pas se former dans des environnements intracellulaires réducteurs (Wörn et Plückthun, 2001). Par ailleurs, retirer des domaines sans implication directe dans la reconnaissance de l'antigène peut faire décliner l'affinité (Holliger et Hudson, 2005). Plus généralement, les anticorps sont souvent limités à des conditions d'utilisation proches de leur environnement physiologique et sont rapidement inactivés dans des conditions induisant un stress important (milieu acide, présence de protéases ou température élevée). Toutes ces limitations ont stimulé un effort de recherche visant à l'optimisation des propriétés des anticorps (moins complexes, plus petits et robustes), tout en conservant leur capacité à fixer spécifiquement un antigène. Différents fragments d'anticorps ont ainsi été générés allant du simple fragment variable à chaîne unique (scFv), au minibody ou encore au diabody (Figure 2 et Figure 3; Skottrup, 2010).

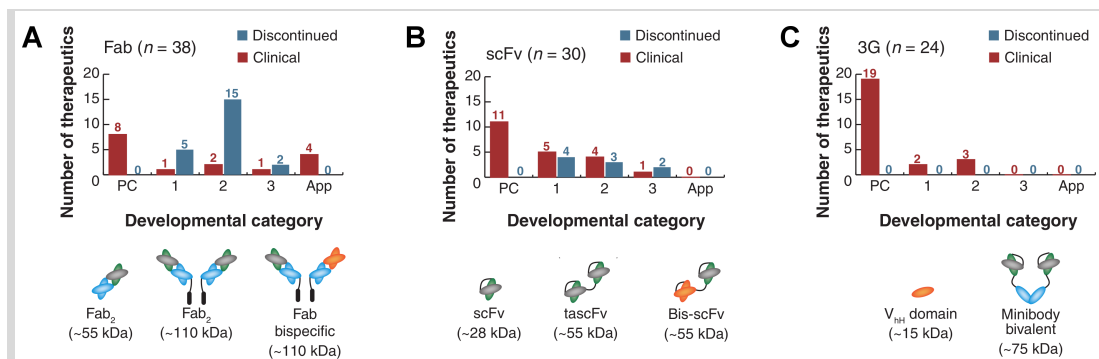


Figure 2: Différentes classes de fragments d'anticorps thérapeutiques en recherche préclinique et développement clinique. PC, préclinique; App, approuvé; 3G, 3ème génération; tascFv, tandem single chain variable fragment; Bis, bispecific. Adapté d'après Nelson et Reichert, 2009.

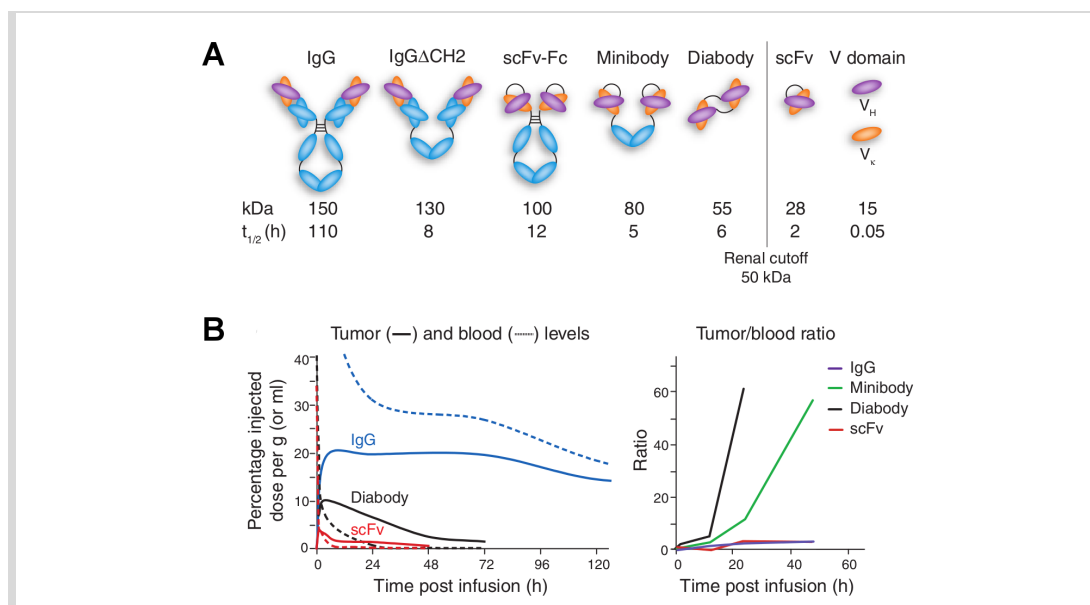


Figure 3 : Importance de la taille des anticorps pour leur pharmacocinétique et leur biodistribution. A) Représentations de différents formats d'anticorps utilisés en imagerie in vivo, ainsi que leur poids moléculaire (kDa) et demi-vie sérique (phase θ). B) Biodistribution de différents formats de mAb dans deux modèles de xénogreffes murines. A gauche : taux d'une IgG, d'un diabody et d'un fragment scFv anti-HER2 dans la tumeur (lignes pleines) ou le sang (lignes pointillées) en fonction du temps après infusion, dans des souris SCID (severe combined immune deficient) porteuses de tumeurs solides sous-cutanées SK-OV-3. A droite : Ratio tumeur/sang d'IgG, minibody, diabody et fragment scFv anti-CEA en fonction du temps après infusion, dans des souris athymiques avec xénogreffe de carcinome colique. Ces données de biodistribution indiquent un meilleur ciblage de la tumeur par les diabodies (assimilation tumorale et clairance sanguine rapide) en comparaison avec les immunoglobulines (en plus grande quantité dans le sang) et les scFv (plus faible présence tumorale). Adapté d'après Holliger et Hudson, 2005.

En parallèle du travail d'ingénierie d'anticorps, différentes charpentes aux propriétés de reconnaissance moléculaire alternatives aux anticorps ont été proposées, parmi lesquelles les Adnectines (Ramamurthy et al., 2012), les Affibodies (J. Li et al., 2010), les Anticalins (Schönfeld et al., 2009), les DARPin (Boersma et Plückthun, 2011) ou les Nanofitines (Kreihenbrink et al., 2008). Les études cliniques les mettant à profit sont pour la plupart encore en cours (**Tableau 1**) mais devraient confirmer le fort potentiel de ces nouvelles charpentes, et possiblement certaines de leurs limites, en tant que médicaments. A ce jour, plus de 50 de ces nouvelles charpentes ont été décrites (Binz et al., 2005; Grönwall et Ståhl, 2009; Hey et al., 2005; Hosse et al., 2006; Nygren, 2008; Skerra, 2007; Škrlec et al., 2015), mettant ainsi en avant leurs caractéristiques communes. Généralement, ce sont des protéines de faible poids moléculaire (moins de 100 acides aminés) avec une structure rigide composée d'une chaîne polypeptidique unique (**Figure 4**) et fixant naturellement une autre biomolécule. Leur cœur très structuré autorise la création de domaines variables permettant des insertions, délétions ou substitutions sans altérer la conformation

globale de la protéine (Hussain et al., 2012). Ces domaines variables peuvent théoriquement générer une interface pour n'importe quelle cible, de façon similaire aux paratopes des anticorps.

Tableau 1: Preuve de concept de charpentes alternatives aux anticorps.

Name	Scaffold or format	Developer or licensee	Parent protein structure	Clinical trial phase	Disease	Target
Ecallantide (Kalbitor/DX88)	Kunitz domain	Dyax	Human lipoprotein-associated coagulation inhibitor (LACI)	FDA approved (December 2009)	Hereditary angioedema	Kallikrein inhibitor
TRU-015	SMIP	Trubion/Pfizer	Various origin and length	Phase IIb	NHL	CD20
Dom-0200/ART621	Domain antibody	Domantis (now GlaxoSmithKline)/Cephalon	VH or VL antibody domain; 100–130 amino acids	Phase II	Rheumatoid arthritis and psoriasis	TNF
MT103	BiTE	Micromet	scFv–scFv; 200–260 amino acids	Phase II	ALL	CD19 and CD3
				Phase I	NHL	
Angiocept (BMS-844203/CT-322)	Adnectin	Adnexus (owned by Bristol-Myers Squibb)	10th FN3 domain of fibronectin; 94 amino acids	Phase II	Colorectal cancer, NSCLC and glioblastoma	VEGFR2
ALX-0081	Nanobody	Ablynx	VHH; ~100 amino acids	Phase II	ACS and TTP	vWF
ESBA105	Stable scFv	ESBATEch/Alcon	scFv with hyperstable properties	Phase II	Uveitis	TNF
AMG-220 (C326)	Avimer	Avidia (owned by Amgen)	Domain A of LDL receptor; a repeating motif of ~35 amino acids	Phase I	Crohn's disease	IL-6
MT110	BiTE	Micromet	scFv–scFv; ~500 amino acids	Phase I	Lung and gastric cancers	EPCAM and CD3
ABY-002	Affibody	Affibody	Z domain of protein A from <i>Staphylococcus aureus</i> ; 58 amino acids	Phase I	Breast cancer imaging	HER2
MP0112	DARPin	Molecular Partners	Ankyrin repeat proteins; 67 amino acids plus a repeating motif of 33 amino acids	Phase I	Ophthalmological diseases	VEGF
PRS-050 (Angical)	Anticalin	Pieris	Lipocalin; 160–180 amino acids	Phase I starts early 2010	Solid tumours	VEGF

ACS, acute coronary syndrome; ALL, acute lymphoblastic lymphoma; BiTE, bispecific T cell engager; DARPin, designed ankyrin repeat protein; EPCAM, epithelial cell adhesion molecule; FDA, United States Food and Drug Administration; HER2, human epidermal growth factor receptor 2; IL, interleukin; LDL, low-density lipoprotein; NHL, non Hodgkin's lymphoma; NSCLC, non-small-cell lung carcinoma; R, receptor; scFv: single-chain variable domain antibody fragment; SMIP, small modular immunopharmaceutical; TNF, tumour necrosis factor; TTP, thrombotic thrombocytopenic purpura; VEGF, vascular endothelial growth factor; V_H, heavy chain variable domain; VHH, heavy chain variable domain (in camelids); V_L, light chain variable domain; vWF, von Willebrand factor. D'après Beck et al., 2010.

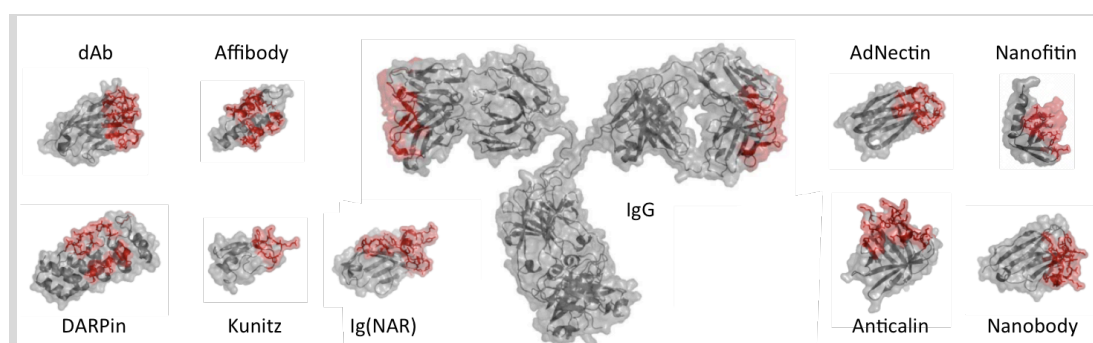


Figure 4: Structures d'anticorps, de fragments d'anticorps et de protéines d'affinités. Les domaines figurés en rouge représentent la surface d'interaction généralement modifiée sur la charpente. Adapté d'après Fiedler et Skerra, 2008.

Du point de vue fonctionnel, ces nouvelles protéines d'affinité peuvent se présenter comme des analogues aux anticorps monoclonaux (**Tableau 2**): un site de fixation homogène; une haute spécificité et une forte affinité dans la gamme du nM au pM (*Gebauer et Skerra, 2009*); la capacité d'inhiber une interaction entre biomolécules (*Mouratou et al., 2007*); une variabilité de séquences, de chaînes latérales, et de conformation du squelette carboné tout en conservant des structures canoniques.

Tableau 2: Comparaison synthétique entre les anticorps et les protéines d'affinité.

Characteristic	Antibody	Binding protein
Size (kDa)	~150–160	<30 (<i>Sac7d</i> : 7 kDa)
Selection	In vivo	In vitro
Production	Animal or recombinant	Recombinant
Post-selection modifications	Possible, but heterogeneous products	Possible, can be designed for homogeneous products
Stability	Several weeks at 4°C Aggregate > 45°C	Variable (<i>Sac7d</i> : pH 0–12, natural habitat at 85°C)
Binding site	Monoclonal: homogeneous; polyclonal: heterogeneous	Homogeneous
Target molecules	Mainly immunogenic macromolecules	Macromolecules and low-molecular-mass molecules
In vivo half-lives	Days to weeks	n.a.
Application conditions	Physiological	Physiological and non-physiological
Therapeutics/treatment	Commercially available	Reported in literature
Targeted drug delivery	Commercially available	Reported in literature
Molecular imaging	Commercially available	Reported in literature
Diagnostics	Commercially available	Reported in literature
Affinity purification	Reported in literature	Commercially available
Biosensors	-	Reported in literature

L'accent est porté sur leurs caractéristiques et les applications auxquelles ils conviennent. n.a., not applicable. Adapté d'après Ruigrok et al., 2011.

Les charpentes alternatives présentent par ailleurs plusieurs avantages par rapport aux anticorps et leurs fragments recombinants. Plus stables, elles peuvent résister à une plus grande gamme de conditions physicochimiques, ce qui facilite les possibilités d'ingénierie par fusion ou par couplage chimique. Souvent dépourvues en cystéine, elles sont également plus faciles à produire et avec de meilleurs rendements. La possibilité d'utiliser un système de production de type bactérie ou levure, associé à des purifications simples (*Aldington et Bonnerjea, 2007*) rend également ces molécules plus abordables. Leur plus petite taille favorise leur pénétration tissulaire (notamment dans les tumeurs solides; *Schmidt et Wittrup, 2009*) et leur capacité

d'accès aux épitopes. Contrairement aux anticorps, elles n'ont pas de fonction effectrice d'immunoglobulines. Pour certaines applications, ces fonctions ne sont pas nécessaires voire même indésirables. Par exemple, une activation inappropriée des cellules exprimant des récepteurs au Fc peut déclencher un relargage massif de cytokines et les effets toxiques qui en découlent. Pour des applications basées sur le déclenchement de voies de signalisation de l'immunité, il est toutefois possible de rétablir une partie des fonctions propres aux immunoglobulines (notamment par couplage de fragment Fc, d'interacteur anti-CD3, etc.).

Finalement, ces charpentes semblent pouvoir s'ouvrir à un champ d'application plus large que celui des anticorps (**Figure 5**).

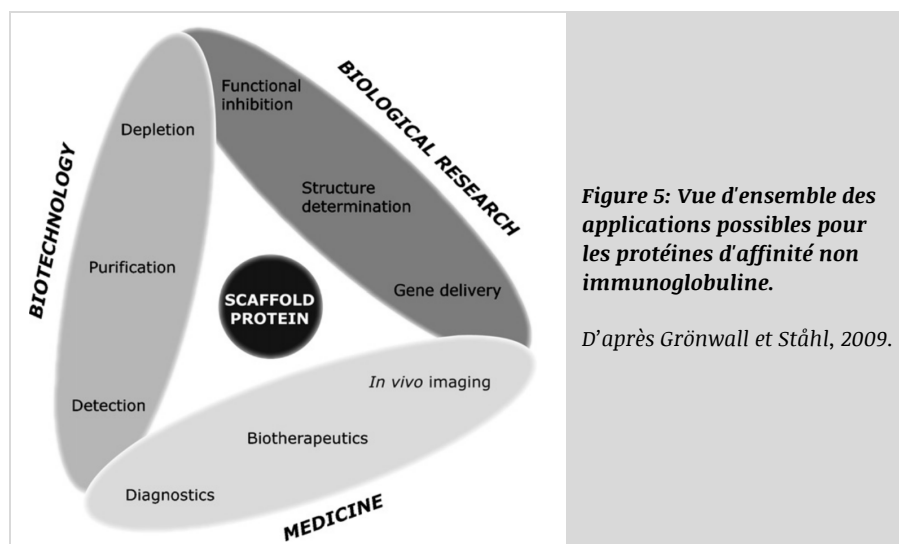


Figure 5: Vue d'ensemble des applications possibles pour les protéines d'affinité non immunoglobuline.

D'après Grönwall et Ståhl, 2009.

I.2. Ingénierie des protéines et apports des outils de modélisation

I.2.1. Optimisations au service du développement de médicaments

Après avoir identifié les candidats les plus spécifiques et affins pour leur cible, il est nécessaire de valider, et souvent optimiser, la compatibilité entre les propriétés des protéines d'affinité et le cadre de leur fabrication puis de leur utilisation, en particulier pour une application thérapeutique. Ceci peut impliquer des améliorations relatives à leur reconnaissance moléculaire (inhérentes à la spécificité et aux constantes d'affinité en particulier), mais peut aussi être lié à leurs paramètres pharmacocinétiques. Ces derniers, à valider *in vitro* et *in vivo*, se traduisent notamment par des études de distribution et de demi-vie, en lien étroit avec les profils de stabilité, de solubilité et d'immunogénicité. Parmi les premières optimisations de protéines

recombinantes, une des approches consistait à greffer des groupements fonctionnels à la surface des protéines pour répondre à une majorité de ces problématiques. Nous pouvons citer en particulier l'exemple de la PEGylation (*Abuchowski et al., 1977*), correspondant à la conjugaison de groupement polyéthylène glycol (PEG). Les protéines ainsi modifiées bénéficient d'une augmentation de taille réduisant la clairance par élimination rénale, d'une diminution de l'accessibilité aux enzymes protéolytiques qui les dégraderaient, d'une exposition réduite aux anticorps avec pour effet de diminuer leur immunogénicité, et d'une augmentation de solubilité grâce à l'ajout de groupement hydrophiles. Néanmoins, cette approche ne présente pas que des avantages puisque la réduction d'accessibilité peut également concerner l'interaction avec la cible de la protéine PEGylée. De plus, la PEGylation sur des sites préférentiels mais non exclusifs peut aboutir à des produits relativement hétérogènes ayant pour effet la réduction de l'activité spécifique des protéines modifiées, suggérant ainsi la nécessité de rationaliser les sites de fonctionnalisation (*Yu Sen Wang et al., 2002; Mei et al., 2010*). Les méthodes de fonctionnalisation chimique spécifique ont d'ailleurs été complétées par des approches de design computationnel des sites de greffe des polymères (positions des résidus conjugués et taille des polymères à greffer) à partir d'ensemble de coordonnées spatiales des protéines d'intérêt (*Desjarlais et al., 2004*).

D'autres méthodes plus spécifiques ont depuis été développées pour s'adresser à chaque propriété des protéines à améliorer, dont les optimisations peuvent être guidées par une ingénierie rationnelle assistée par modélisation moléculaire. Initialement développés pour l'analyse de données expérimentales (dont un espace de séquence de gènes et de protéines volumineux), les outils bioinformatiques se sont spécialisés en diverses branches dont la modélisation moléculaire, incluant la visualisation puis la prédiction de structures et de fonctions des protéines. Bien que limités à notre capacité à retranscrire des données expérimentales en modèles théoriques, ces aspects computationnels peuvent apporter des gains significatifs à l'optimisation de protéines, en particulier vis-à-vis de leur reconnaissance moléculaire mais également par rapport à leur profil de produit cible (*Target Product Profile*, ou TPP, étant associé à des exigences en matière de stabilité, toxicité, solubilité, immunogénicité et demi-vie dans le

cadre d'un produit thérapeutique, en fonction du site d'action et du type d'activité souhaités), comme nous allons l'illustrer dans cette section.

I.2.2. Stabilité et solubilité

Les notions de stabilité et de solubilité des protéines sont fondamentales pour que ces dernières puissent exercer leur activité et répondre aux exigences d'un produit thérapeutique, d'autant plus dans un contexte *in vivo* impliquant un environnement complexe pouvant exposer les molécules thérapeutiques à des stress variés tels que des variations de pH ou une exposition à des enzymes protéolytiques. Un criblage expérimental de ces conditions est possible (via des tests combinatoires de stabilité en présence de différents tampons et protéases à différentes températures par exemple) mais demeure contraignant et nécessite un débit élevé.

Des approches prédictives ont donc été mises en place pour optimiser les interactions protéine-solvant, limiter l'agrégation et améliorer la compaction des protéines, tout en conservant une activité optimale. Par exemple, une analyse de corrélation entre séquence et taux de solubilité de protéines abondantes (de la famille des albumines sériques et des myoblobines) et de leurs paralogues moins abondants a récemment suggéré que l'augmentation du ratio lysine:arginine pouvait accroître la solubilité des protéines et/ou leur niveau d'expression (Warwicker *et al.*, 2014). Par ailleurs, des études de solubilité et de thermostabilité ont également été menées à l'aide de données structurales. Ainsi, des fragments d'anticorps thermorésistants ont été générés par design à partir de modèles d'homologie puis de calculs d'énergie (Miklos *et al.*, 2012). Tout en conservant une fixation spécifique et affine (avec une affinité multipliée par 30), les fragments scFv superchargés générés ont démontré une résistance accrue à la température grâce à une capacité de repliement améliorée. L'utilisation de données structurales pour la thermostabilisation rationnelle de protéine a également été illustrée lors d'une démonstration appliquée à l'apoflavodoxine, précédemment décrite comme difficile à stabiliser (Lamazares *et al.*, 2015). En identifiant les régions les plus instables de la protéine, les auteurs ont ainsi prédit des mutations menant à l'obtention de variants avec une température de fusion de 75°C (soit 32°C de plus que la protéine sauvage). Ces exemples récents illustrent l'apport des prédictions

computationnelles qui peuvent être appliquées à l'optimisation des profils de stabilité et de solubilité de protéines, en s'appuyant sur des profils de séquences ou sur l'agencement spatial des acides aminés dans des études de modélisation par homologie ou *de novo*.

I.2.3. Demi-vie

La demi-vie des protéines est fortement liée à leur stabilité ainsi qu'à leur solubilité, mais est également dépendante de la taille des protéines qui détermine leur clairance (Bocci, 1990). De ce fait, de petites protéines bénéficient en règle générale d'une clairance élevée, ce qui peut constituer un atout essentiel pour des applications telles que l'imagerie médicale. A l'inverse, ce paramètre représente un inconvénient majeur pour les protéines de plus petite taille dans le cadre d'une utilisation thérapeutique avec administration systémique.

Plutôt que d'augmenter la taille des protéines dont la demi-vie doit être étendue, une stratégie consiste à ajouter un groupement fonctionnel se liant à des protéines circulantes ou permettant le recyclage plutôt que l'élimination (voir Kontermann, 2011 pour une revue des stratégies récentes). Un exemple répandu parmi les modules d'extension de demi-vie fusionnés à des protéines recombinantes thérapeutiques est l'utilisation de domaines de fixation à l'albumine sérique (présente en abondance et possédant une longue demi-vie), allant de courts peptides à des domaines de fixation complets (Dennis *et al.*, 2002; Holt *et al.*, 2008). Ce type d'approche a permis de compenser efficacement la petite taille de charpentes alternatives aux anticorps, tout en étant facilité par leur ingénierie aisée. Enfin, une stratégie innovante pour diminuer la clairance de protéines d'affinité pourrait être basée sur le transfert de leur site de fixation sur une charpente humaine circulante (stratégie de greffe), plutôt que d'y être liée (par conjugaison ou fusion génétique), à partir d'analyse de données structurales comme nous le détaillerons dans la section "1.3. Transfert de fonction et stratégie d'humanisation".

I.2.4. Immunogénicité

L'humanisation de protéine par transfert d'un domaine protéique vers une charpente humaine est par ailleurs communément adoptée pour limiter l'introduction de peptides étrangers au soi lors de l'administration de protéines thérapeutiques, et ainsi réduire le risque de réaction

immunitaire induite par une réponse des cellules T CD4+ (Baker et al., 2010). En effet, des peptides exogènes présentés aux cellules immunitaires à la surface des complexes majeurs d'histocompatibilité de type II avec une forte affinité sont connus pour activer la réponse immunitaire s'ils sont identifiés comme étrangers au soi. La suppression des peptides immunogènes a été largement employée avec les anticorps, via l'utilisation d'anticorps humains ou humanisés grâce aux homologies entre les immunoglobulines de différentes espèces. Cependant, il n'existe pas systématiquement d'homologues humains dans le cas des charpentes non immunoglobulines, et ceux-ci ne sont pas nécessairement circulants s'ils existent, rendant d'autres approches nécessaires.

L'arsenal des méthodes de détection et prédiction du risque d'immunogénicité est actuellement complété par des techniques *in silico*, *in vitro* et *in vivo* (Deehan et al., 2015). Brièvement, le profil immunogénique d'une protéine provient de paramètres variés, généraux ou spécifiques au patient (Figure 6A), dont une partie peut être résolue par l'ingénierie rationnelle des protéines. Il est notamment possible d'atténuer efficacement le déclenchement de l'immunité via une réduction de la dégradation et de l'agrégation du médicament (comme discuté précédemment). Il est également possible d'identifier et prédire les protéines les moins immunogènes ou les régions à muter (Figure 6B). Ce type de désimmunisation a d'ailleurs été récemment approfondi en intégrant des informations structurales, avec une méthode de suppression des épitopes T intégrée à la suite de modélisation Rosetta (Choi et al., 2013).

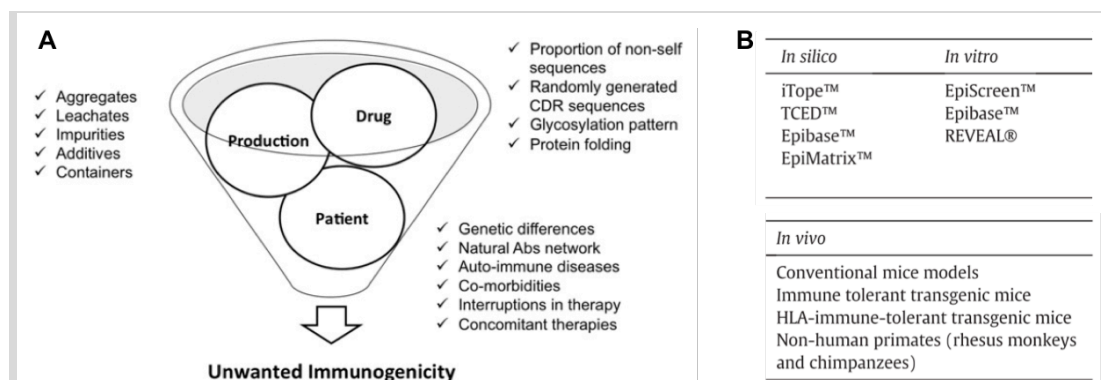


Figure 6: Paramètres impactant l'immunogénicité des protéines. A) Paramètres menant à une immunogénicité non désirée. Dans le cas des Nanofitines, les régions randomisées peuvent générer des peptides immunogènes ou non-immunogènes, de la même façon que les CDR des anticorps. Les motifs de glycosylation ne concernent pas les Nanofitines, non glycosylées du fait de leur production en bactéries. B) Résumé des modèles et techniques disponibles *in silico*, *in vitro* et *in vivo*. Adapté d'après Deehan et al., 2015.

I.2.5. Repliement et fonction

Les méthodes récemment développées font de plus en plus fréquemment appel à une caractérisation structurale pour optimiser les protéines recombinantes, de leur stabilité à leur immunogénicité. D'une part, ceci est lié à la disponibilité grandissante de structures tridimensionnelles disponibles dans la banque de données sur les protéines du *Research Collaboratory for Structural Bioinformatics* (Berman et al., 2000). D'autre part, ceci témoigne de l'amélioration des méthodes *in silico* relatives à l'analyse et la prédiction du repliement protéines.

Diverses stratégies ont été envisagées au cours des 30 dernières années, en considérant les protéines du niveau atomique jusqu'à celui de blocs structuraux (dont une approche d'alphabets structuraux a été récemment intégrée au sein de l'UFIP; Mahajan et al., 2015). L'avancement de l'ensemble de ces stratégies a ouvert la voie aux premiers designs *de novo* de protéines, dont des exemples marquants sont la création de petites protéines en hélices (Hecht et al., 1990), en doigt de zinc (Dahiyat et Mayo, 1997), en feuillet beta (Kortemme et al., 1998), ou globulaires (Kuhlman et al., 2003). Parmi ces exemples de prédictions avérées de structures, nous retrouvons des contributeurs du logiciel de modélisation Rosetta, dont les succès se sont notamment illustrés lors des différentes éditions du test de prédiction structurale CASP (*Critical Assessment of Techniques for Protein Structure Prediction*), grâce à des combinaisons d'approches *de novo* et par homologie (Raman et al., 2009; Taylor et al., 2014). Nous détaillerons également notre intérêt pour cette suite logicielle dans la section "I.4. Approches computationnelles de design rationnel", car ses résultats prometteurs ont été récemment étendus à la prédiction d'interface protéine-protéine et au ciblage précis d'épitopes définis.

I.2.6. Applications aux Nanofitines, protéines d'affinité développées à Affilogic

Les travaux présentés dans le présent mémoire ont été réalisés dans le cadre d'une thèse en convention industrielle de formation par la recherche (CIFRE, convention N°2012/1337) co-financée par le Fonds européen de développement régional (FEDER, convention 2013/FEDER/254 référence 39245) dans l'Unité Fonctionnalité et Ingénierie des Protéines (UFIP), laboratoire UMR CNRS 6286,

et au sein d'Affilogic, société nantaise créée en Février 2010. La spécialité de cette entreprise est le développement et l'optimisation de Nanofitines, autour desquelles gravitent les travaux présentés dans notre étude. Ces protéines d'affinité, notamment en phase de développement pharmaceutique pour la création d'agents neutralisants à usage thérapeutique, sont classées parmi les charpentes alternatives aux anticorps. En effet, les Nanofitines sont dérivées de Sac7d, une protéine à structure *OB-fold* en tonneau beta incomplet à 5 brins (Murzin, 1993) extrêmement stable (notamment à la chaleur et au pH) issue de l'archée hyperthermophile *Sulfolobus acidocaldarius*. Elle est composée de 66 acides aminés (environ 7 kDa) dont aucune cystéine, et a la capacité de fixer naturellement l'ADN double brin (Edmondson et Shriver, 2001). Des banques de mutants de Sac7d ont été générées via l'introduction de mutations aléatoires par substitution à la surface de la protéine, notamment sur les résidus du feuillet beta normalement impliqués dans la fixation de l'ADN (Mouratou et al., 2012). Les variants de Sac7d sélectionnés par la technique du *ribosome display* contre une cible donnée sont appelés Nanofitines (Figure 7).

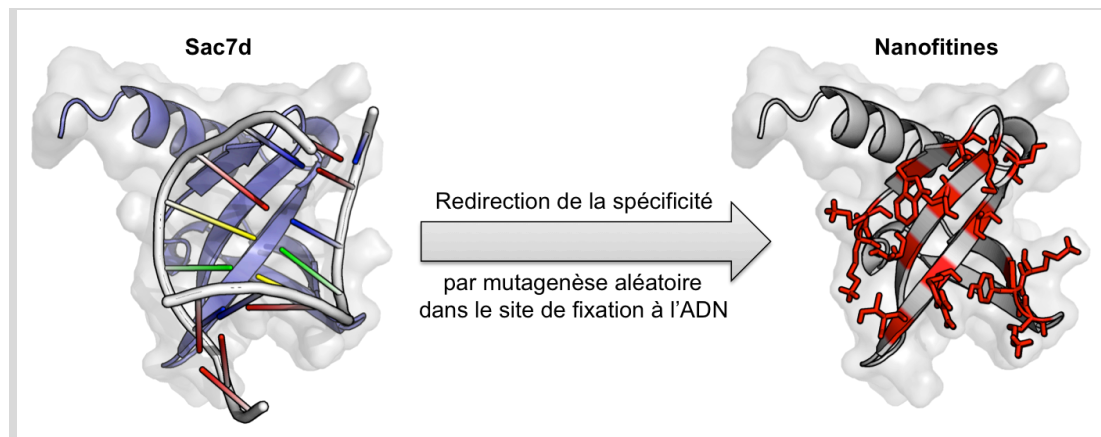


Figure 7: Représentation schématique de la génération de Nanofitines à partir de la charpente de Sac7d. La charpente de Sac7d sauvage (en violet) est capable de se lier à un duplex d'ADN (squelette phospho-sucrose en gris, avec représentation des bases azotées en bâtonnets rouge, jaune, vert et bleu). Les résidus impliqués dans la fixation de l'ADN (en rouge) sont randomisés pour sélectionner des Nanofitines avec une spécificité redirigée envers une nouvelle cible. Identifiant PDB de la structure de référence de Sac7d: 1AZP.

Grâce à cette technologie, Affilogic possède aujourd'hui un catalogue de Nanofitines, dont des agents bloquants, dirigés contre diverses cibles (Prel, 2015). L'exploitation des propriétés de robustesse, du fort potentiel de pénétration tissulaire, et des couplages ou fusions aux Nanofitines a notamment été étudiée dans les domaines de l'inflammation, de l'oncologie, de la neurologie, de l'ophtalmologie ou des maladies infectieuses. Dans ces stratégies, les avantages

des Nanofitines au regard des inconvénients de stabilité et de taille des anticorps ont notamment permis d'explorer des voies d'administration innovantes et moins invasives, telles que les administrations orale et cutanée. Pour orienter la découverte de l'ensemble de ces Nanofitines vers les protéines les plus efficaces, la sélection par *ribosome display* a été complétée par la mise en place de conditions de criblage adaptées à chaque application visée. Cet effort, intensifié pour les stratégies de sélection les plus élaborées, pourrait être soulagé par une rationalisation apportée par des outils de modélisation, ce qui représentera le sujet central de ce manuscrit.

Sur les aspects de solubilité, de stabilité et de repliement, les connaissances accumulées autour des Nanofitines dans la littérature¹ et par Affilogic laissent présager des propriétés favorables de la charpente (donnant généralement accès à des variants solubles et extrêmement stables au pH, à la température et aux molécules chimiques malgré l'introduction de mutations). Nous avons donc proposé une approche d'humanisation des Nanofitines pour anticiper les études liées aux profils de demi-vie et d'immunogénicité des Nanofitines, très récemment initiées au moment du démarrage des travaux décrits dans ce manuscrit. Par ailleurs, le procédé de sélection *in vitro* des Nanofitines donne accès à des variants d'affinité généralement nanomolaire, avec des stratégies de maturation d'affinité additionnelle qui ont déjà été réalisées avec succès. La composante d'affinité de la reconnaissance moléculaire est donc efficacement contrôlée, mais la maîtrise précise des épitopes ciblés demande encore d'intenses efforts de criblage pour les projets les plus fins. Nous avons donc également cherché à rationaliser l'ingénierie de la surface de fixation des Nanofitines envers leurs cibles par design *de novo*.

Nous avons mis en place la preuve de concept de chacune de ces approches (humanisation et design *de novo*) en ciblant la protéine de fluorescence verte (GFP). En plus d'être une cible largement documentée (jusqu'à la publication de structures en complexe), la GFP représente un outil d'intérêt biotechnologique et bénéficie de plusieurs protéines d'affinité spécifiques dirigées contre elle (*AntibodyRegistry*, 2015; *Brauchle et al.*, 2014; *Guellouz et al.*, 2013; *Koide et al.*, 2012;

¹ Une liste de publications centrées sur la charpente des Nanofitines: *Béhar et al.*, 2013, 2014; *Buddelmeijer et al.*, 2009; *Cinier et al.*, 2009, 2012; *Correa et al.*, 2014; *Gera et al.*, 2012, 2011; *Hussain et al.*, 2013; *Krehenbrink et al.*, 2008; *J. Il Lee et al.*, 2010; *M. Li et al.*, 2014; *Miranda et al.*, 2011; *Mouratou et al.*, 2007, 2012; *Pacheco et al.*, 2014; *Ppyun et al.*, 2012; *Varga et al.*, 2014; *Yan Wang et al.*, 2004; *Wu et al.*, 2005; *Yang et Wang*, 2004.

Rothbauer et al., 2008). Par ailleurs, son étude était facilitée par les outils disponibles au laboratoire, en particulier pour sa production sous forme recombinante (à partir de plasmides à notre disposition) et son contrôle qualité (chromatographie, électrophorèse, absorbance, fluorescence, reconnaissance moléculaire).

Après avoir décrit plus précisément les contextes respectifs des approches complémentaires du transfert de charpente protéique et du design *de novo* de protéine, nous consacrerons un chapitre à chaque stratégie pour discuter des résultats générés lors de notre preuve de concept.

I.3. Transfert de fonction et stratégie d'humanisation

I.3.1. Pourquoi humaniser les protéines à usage thérapeutique ?

Avant de détailler comment une fonction protéique peut être transférée à une autre protéine par greffe d'acides aminés fonctionnels, nous allons aborder le but de cette stratégie dans le cadre de l'humanisation de protéines destinées à une application thérapeutique. Les Nanofitines sont des variants artificiellement dérivés de la protéine Sac7d, dont l'origine est archéobactérienne. De ce fait, les séquences en acides aminés qui les composent ne correspondent pas à celles de protéines humaines. Au démarrage des travaux menés pendant la thèse décrite dans ce manuscrit, les rares études d'immunogénicité approfondies portées à notre connaissance (avec un nombre suffisant de variants pour pouvoir estimer le potentiel immunogène de la charpente protéique ou des régions randomisées) ne laissaient pas l'opportunité de dresser des profils d'immunogénicité fiables des Nanofitines, laissant des incertitudes quant à leur capacité à être utilisées en thérapeutique (en particulier lors d'administrations systémiques répétées) sans déclencher de réaction immunitaire non désirée. L'origine non-humaine des Nanofitines était donc souvent associée à un risque élevé d'immunogénicité dans l'esprit collectif, indépendamment d'autres paramètres désormais bien identifiés comme l'agrégation et la solubilité de la protéine en question. Les Nanofitines présentent d'ailleurs des caractéristiques relativement favorables du point de vue de ces derniers paramètres, indépendantes de l'organisme d'origine de la charpente protéique, car ce sont de petites protéines, globulaires et très solubles, ce qui limiterait le risque d'immunogénicité induite par une agrégation.

La problématique liée à l'administration de protéines exogènes a déjà été intensément étudiée pendant les dernières décennies dans le développement d'anticorps administrés à des patients. Initialement générés par immunisation de rongeurs, l'utilisation des anticorps non-humains a soulevé des questions quant au déclenchement de réactions immunitaires humaines chez le patient, malgré les homologues entre les équivalents murin et humain. En réponse à ce risque, la stratégie d'humanisation a rapidement été adoptée et appliquée généralement aux anticorps (Hurle et Gross, 1994). Grâce à cette approche, le nombre d'acides aminés reconnus comme étrangers par le système immunitaire est limité, diminuant ainsi le risque de réaction dirigée contre la protéine administrée. En pratique, les stratégies d'humanisation se résument à modifier une charpente protéique humaine par transfert des résidus responsables de l'activité d'une protéine d'origine non-humaine (issue d'animaux, de plantes, de micro-organismes, ou générée *de novo*).

Comme nous le détaillerons à travers une série d'exemples d'intérêt, de tels transferts d'activité inter-protéiques n'ont pas seulement été effectués sur des anticorps (Vita, 1997). En effet, des succès de greffes de résidus entre diverses charpentes protéiques ont été illustrés dans la littérature, entre homologues structuraux mais également vers des charpentes alternatives adaptées à la présentation du site transféré. Souvent cantonnés à un nombre restreint d'acides aminés, ces transferts peuvent s'étendre à plusieurs éléments complets de structure secondaire, avec des motifs continus ou discontinus, pour induire des changements de spécificité de substrat, de coenzyme ou de capacité de liaison. Par ailleurs, la possibilité de transférer une activité biologique entre protéines distinctes pourrait ouvrir la voie à des découvertes intéressantes en matière de propriété intellectuelle, afin d'étendre ou de démarquer la recherche menée au laboratoire via la génération de brevets sur de nouvelles charpentes protéiques.

I.3.2. Succès de greffes protéiques

Dans le cadre de ces travaux de thèse, nous nous sommes intéressés à la greffe des acides aminés fonctionnels localisés sur 3 brins beta de la charpente des Nanofitines, comprenant donc 11 résidus de surface séparés par 7 résidus enfouis, vers une protéine hôte d'origine humaine.

Bien qu'il existe de nombreuses illustrations de transfert de site dans la littérature, que nous détaillerons ci-après, les exemples de greffe d'un motif de plus de 5 résidus demeurent relativement anecdotiques lorsqu'aucune boucle n'est concernée, soulignant ici le défi technique d'une humanisation en une seule étape via le transfert du site de fixation complet d'une Nanofitine.

Afin de réaliser un tour d'horizon des succès de transfert de site (et des limitations qu'ils mettent en avant), nous pouvons classer artificiellement ces exemples en différentes catégories. Tout d'abord, nous aborderons des illustrations de transfert d'activité sans interaction protéine-protéine (coordination de métaux, activités enzymatiques ou structures modèles). Ensuite, nous focaliserons notre attention sur la greffe de motifs aux interfaces inter-protéines en illustrant des transferts de boucles, le mimétisme de ligands naturels puis le transfert de faces complètes de feuillets beta.

I.3.2.1. Sites de coordination d'ions métalliques

Le transfert de site de coordination figure parmi les premiers exemples de greffe de résidus fonctionnels entre protéines. En règle générale, ce type de stratégie implique 3 à 4 résidus dans la formation des liaisons de coordination avec des ions métalliques. Les illustrations les plus connues concernent les domaines de fixation au zinc, avec par exemple le transfert de 9 acides aminés (dont 3 histidines) de l'anhydrase carbonique B sur deux brins beta anti-parallèles d'une toxine de scorpion (*Vita et al., 1995*). Dans cette étude, la protéine correctement exprimée a bénéficié d'une nouvelle activité lui permettant de fixer préférentiellement les ions Cu^{2+} ($K_D = 42 \text{ nM}$), et en moindre mesure Zn^{2+} ($K_D = 5,3 \mu\text{M}$), Cd^{2+} , Mn^{2+} et Ni^{2+} . Le même motif discontinu composé d'histidines a également été transféré sur des charpentes protéiques différentes, en hélices (*Handel et DeGrado, 1990*) ou en tonneau beta (*Müller et Skerra, 1994*). Dans ce dernier exemple, le transfert du site de coordination a été réalisé sur des lipocalines, montrant qu'un clone greffé du motif poly-histidine pouvait fixer le zinc avec une affinité de 36 nM tout en conservant son activité naturelle de fixation à l'acide rétinoïque, ce qui autorisait ainsi sa purification par IMAC avec des résines de Ni^{2+} (chélaté de façon similaire au Zn^{2+}). D'autres sites

de fixation au zinc ont également été décrits, comme le site de coordination tétraédrique formé par 2 résidus histidine et 2 résidus cystéine retrouvés dans les structures caractéristiques des protéines en doigts de zinc (Regan et Clarke, 1990).

Plus récemment, la greffe d'un site de coordination métallique a démontré pouvoir induire l'adoption d'une conformation en hélice lors de la présence de Fe^{3+} dans le cadre du transfert d'une protéine de la famille des frataxines vers la thioredoxine isolée chez *E. coli* (Vazquez et al., 2015). Cette étude illustre le besoin de tenir compte des mouvements induits lors de la greffe de résidus sur des polypeptides, tout comme dans la description de la méthode *OptGraft* validée sur des exemples préalablement décrits et appliquée avec succès au transfert d'une poche de fixation au calcium (Fazelinia et al., 2009). Pour cela, les auteurs ont évalué la labilité des résidus à greffer et les modifications apportées par leur nouvel environnement, notamment par des calculs d'énergie et en simulant une relaxation du squelette peptidique couplée à une exploration des rotamères du site transféré. La flexibilité du site à greffer d'une protéine à une autre a également été au centre de travaux visant à transférer une boucle de 12 acides aminés de la calmoduline, participant à la formation d'un motif en main EF avec une activité de liaison au calcium (Ye et al., 2003). Grâce à l'insertion de cette boucle à différentes régions du domaine 1 de l'antigène lymphocytaire CD2 de rat (entre des jonctions flexibles poly-glycine), les auteurs ont décrit que la charpente receveuse conservait sa structure tout en formant un complexe 1:1 avec l'ion Ca^{2+} . Du reste, cette stratégie réussie de transfert de boucle a été généralement adoptée pour procurer des activités diverses à des protéines, dont celles que nous aborderons dans les sections suivantes.

1.3.2.2. Activités enzymatiques

Bien qu'il s'agisse d'interactions différentes de celles qui nous intéressent au sein de ce manuscrit, puisque les enzymes interagissent avec leur substrat et/ou co-facteurs via la formation de liaisons covalentes transitoires, il est à noter qu'il est possible de transférer un domaine catalytique à une protéine via une approche de greffe. Nous pouvons notamment citer une étude d'échange de boucles (de 4 à 19 résidus) sur des charpentes α/β_8 issues de différentes enzymes homologues d'une phosphoribosylanthranilate isomérase (Ochoa-Leyva et al., 2009). Par

l'intermédiaire de la fusion avec une protéine rapporteuse, les auteurs ont pu constater que 30% à 90% des variants ainsi générés se sont repliés sous forme soluble, mais ont également rapporté des modifications structurales induites lors de la greffe. Une approche similaire de transfert de boucles (entre 9 et 11 résidus) a aussi montré la possibilité de changer efficacement la spécificité du motif de séquence nucléotidique reconnu pour modifier des enzymes de la famille des cytidine désaminases (Kohli *et al.*, 2009).

I.3.2.3. Modèles de protéines membranaires

Les protéines membranaires, et en particulier les récepteurs couplés aux protéines G (RCPG), représentent un des challenges majeurs de l'étude des protéines ciblées par des molécules thérapeutiques (Filmore, 2004), justifiant un effort dans leur expression, purification et caractérisation structurale comme nous l'évoquerons plus en détails dans la section "I.4.3.2. Résolution de structure". Une approche visant à contourner les difficultés à obtenir ces informations a été récemment proposée par transfert des boucles extracellulaires impliquées dans la fixation du ligand d'un RCPG (le récepteur Y), normalement constitué de 7 hélices transmembranaires (Walser *et al.*, 2012). Les auteurs ont évalué la capacité à mimer les propriétés de ce récepteur en greffant les boucles sur une structure en tonneau beta (sur une protéine soluble et une protéine membranaire), dans le but de proposer un modèle simplifié pour l'étude des interactions avec ses ligands. Malheureusement, les affinités observées sont bien plus faibles que celles obtenues avec le récepteur sauvage, passant du faible nanomolaire au faible micromolaire, mettant en exergue la simplification trop poussée de ce type de modèle.

Ces différents exemples de greffes, appliqués au transfert de site de chélation, de site catalytique enzymatique ou encore aux protéines membranaires telle que les RCPG, mettent en évidence les limitations à surpasser pour prédire facilement, mais surtout précisément, les arrangements permettant un transfert réussi. Ces études ont permis de développer un pan de recherche et des outils de modélisation, bien qu'imparfaits, qui ont également été appliqués à la greffe d'interfaces protéine-protéine. Au cœur de nos préoccupations pour l'humanisation des Nanofitines, le transfert de cette activité de liaison a trouvé son modèle dans la greffe de boucles, comme c'est

classiquement le cas via le transfert de régions déterminant la complémentarité (CDR) des anticorps, ce qui est par ailleurs la première utilisation des immunoglobulines comme charpentes protéiques au sens strict.

I.3.3. Greffes d'interfaces protéine-protéine

I.3.3.1. Transferts de boucles: humanisation classique d'anticorps

Dès la fin des années 1980, les premières greffes de CDR ont démontré la capacité à transférer l'activité de fixation spécifique d'antigène depuis des anticorps murins vers des anticorps humains (**Figure 8**). Par exemple, les CDR de chaîne lourde d'un anticorps murin anti-haptène ont été humanisées pour obtenir des affinités semblables (de l'ordre du micromolaire) avant et après transfert (*Jones et al., 1986*). De même, les 6 CDR d'un anticorps de rat dirigé contre des lymphocytes ont été greffées sur un anticorps humain permettant de conserver une fixation spécifique, mais diminuée d'un facteur 40 d'après les expériences menées en ELISA (*Riechmann et al., 1988*). Toutefois, cette étude a aussi démontré que cette altération pouvait être corrigée par l'introduction de mutations ponctuelles pour améliorer la compaction des chaînes latérales greffées, aboutissant à une perte d'affinité finale d'un facteur 3 seulement. La stratégie d'humanisation d'anticorps par transfert de CDR n'est cependant pas systématiquement réussie telle quelle, puisque des résidus non impliqués directement dans l'établissement d'une interface avec des antigènes se sont révélés primordiaux pour cette reconnaissance moléculaire (*Caldas et al., 2003*).

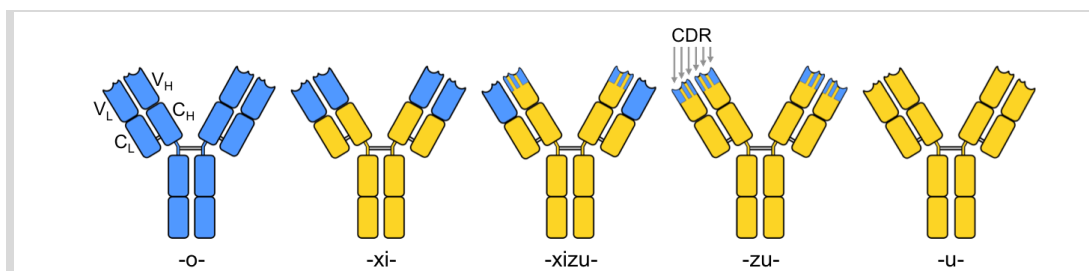


Figure 8: Représentation schématique de constructions intermédiaires entre anticorps murins et humains. Les régions humaines et murines sont figurées, respectivement, en jaune et en bleu. La nomenclature des anticorps est indiqué sous chaque construction: -o- (murin), -xi- (chimérique), -xizu- (chimérique/humanisé), -zu- (humanisé), -u- (humain). Les domaines V_L , C_L , V_H et C_H indiquent, respectivement, les domaines variables des chaînes légères, constants des chaînes légères, variables des chaînes lourdes et constants des chaînes lourdes. Les boucles constituant les CDR sont indiquées par des flèches.

Les boucles transférées dans les nombreux travaux d'humanisation d'anticorps ont donc généralement été greffées sur leur emplacement homologue. Il a également été suggéré que la symétrie des fragments variables à chaîne unique (scFv) des anticorps pourrait permettre d'utiliser la face opposée aux CDR (c'est-à-dire la région orientée vers les domaines constants des scFv, et non les paratopes). Cette théorie originale, proposée lors d'une étude de modélisation menée par Keck et Huston (*Keck et Huston, 1996*), ne semble pas avoir été testée mais plutôt délaissée au profit du transfert sur d'autres charpentes protéiques, homologues ou alternatives aux immunoglobulines, telles que nous le détaillerons ci-après.

I.3.3.2. Transferts de boucles: miniaturisation de charpentes

Avec la découverte de protéines plus petites et stables comme alternatives aux anticorps, le transfert de CDR a été une des stratégies adoptées pour générer des protéines d'affinité mimétiques des anticorps, en se basant sur les informations structurales des charpentes donneuses et receveuses (*Umetsu et al., 2010*). Les travaux menés dans ce contexte ont ainsi mis en avant de nombreux cas d'interchangeabilité entre anticorps et "mini-protéines", identifiant ainsi une large gamme de charpentes utilisables pour développer des interactions protéine-protéine à façon (**Figure 9**).

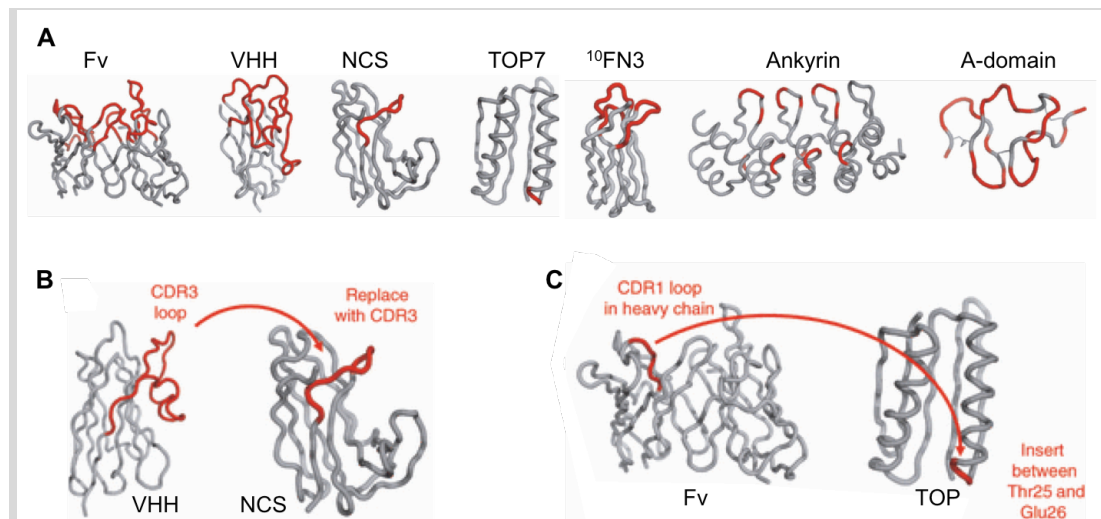


Figure 9: Exemples de charpentes protéiques pour le transfert de boucle. A) Structures de petites charpentes de protéines d'affinité à boucles. Les boucles rouges représentent les régions décrites avec capacité de fixation via transfert de boucle. B) Exemple de transfert de la longue boucle CDR3 d'un anticorps vers la néocarzinostatine. C) Exemple de transfert de la boucle courte CDR1 de la chaîne lourde d'un anticorps vers la charpente TOP. Fv: fragment de la région variable des anticorps. VHH: chaîne lourde variable d'anticorps à chaîne lourde de camélidés. NCS: néocarzinostatine. ¹⁰FN3: 10^{ème} domaine de fibronectine de type III. Ankyrin: Ankyrine. Adapté d'après Umetsu et al., 2010.

Du fait des limitations liées à la charpente immunoglobuline des anticorps, des études précoces ont été menées sur des protéines de la famille des toxines de scorpion, dont certains représentants ont l'avantage de présenter la majorité des éléments structuraux communément retrouvés dans les protéines (hélice alpha, feuillet beta, coude beta et boucles), sur une chaîne unique d'environ 40 résidus stabilisée par des ponts disulfures. Par exemple, la charybdotoxine a été greffée de boucles (et d'une portion des deux brins antiparallèles qui les entourent) pour lui procurer une activité de mimétique du cluster de différenciation CD4 humain, ou de mimétique du curare (Vita *et al.*, 1998). Pour cela, la région du CD4 humain similaire à une CDR2 a pu être efficacement transférée, procurant à la toxine la capacité de se lier à la protéine gp120 du virus de l'immunodéficience humaine (VIH). Dans cette même étude, une boucle issue d'une toxine de serpent a également été greffée sur la toxine de scorpion, lui permettant de fixer le récepteur à l'acétylcholine.

La greffe de boucle d'anticorps a été exécutée sur diverses charpentes alternatives, dont les VHH (chaînes lourdes variables d'anticorps à chaîne lourde de camélidés) ou les minibodies (Martin *et al.*, 1999). Dans les deux cas, la greffe de la capacité de fixation à la protéase NS3 a été décrite, générant des protéines inhibitrices avec une constante d'inhibition (K_i) similaire à celle de l'anticorps donneur (de l'ordre du micromolaire). Cette étude a également mis en avant la variabilité des résultats en fonction de la nature de la boucle à remplacer lors du transfert, qui peut déterminer le succès de l'approche ou produire des protéines insolubles. D'autres homologues structuraux du domaine variable des immunoglobulines ont été validés pour le transfert de CDR, dont la néocarzinostatine, une protéine antibiotique à action anti-tumorale composée de 7 brins beta reliés par des boucles (Nicaise *et al.*, 2004). Après greffe de la CDR3 d'un VHH anti-lysozyme vers son site homologue à la surface de la néocarzinostatine, le variant généré a pu être exprimé de façon stable avec une structure semblable à la protéine sauvage, tout en gagnant une activité de fixation au lysozyme. Toutefois, les auteurs ont rapporté que l'affinité de 20 nM avant transfert a été réduite au micromolaire après transfert, ce qui proviendrait de contraintes structurales non conservées lors de cette approche (comme l'établissement d'un pont disulfure entre les CDR1 et CDR3, absent sur la charpente receveuse).

Il est aussi à noter que les charpentes receveuses des boucles greffées n'ont pas toujours été limitées à une structure unique et déterminée à l'avance. Par exemple, des travaux ont mis à profit un criblage de la banque de données sur les protéines du *Research Collaboratory for Structural Bioinformatics (Protein Data Bank, ou PDB)* pour identifier des protéines humaines receveuses (Inoue et al., 2011). Cette stratégie a mené au transfert d'une boucle de CDR de VHH (d'affinité nanomolaire pour le lysozyme) vers une ubiquitine humaine, aboutissant à un K_D supérieur à 230 μM envers le lysozyme. Les auteurs ont imputé cet effondrement d'affinité à une déstabilisation de la région autour du site greffé (observée par échange hydrogène-deutérium en spectrométrie de masse), résultant en une orientation mal conservée de la CDR3. A titre de comparaison, le même transfert opéré sur un autre VHH (représentant 9 résidus de différence), avait généré des variants anti-lysozyme avec des K_D compris entre 2,8 et 7,3 μM , soulignant les fortes pertes d'affinité à anticiper lors de ce type de transfert. Par ailleurs, la même équipe a par la suite optimisé les résultats du transfert de CDR3 inter-VHH en réduisant la perte d'affinité par un facteur 20 grâce à une approche de design *in silico* (Inoue et al., 2013).

A mesure du développement des stratégies de greffe de résidus fonctionnels, les approches se sont finalement diversifiées, proposant d'intervertir des charpentes hôtes homologues tout en conservant l'activité apportée par des boucles exposées en surface, ou en générant de nouvelles activité à partir de ligand naturels à mimer. On retrouve donc dans la littérature des travaux s'étendant de l'ingénierie d'un nanobody humanisé "universel", décrit pour recevoir les CDR greffées de VHH (Vincke et al., 2009), à l'extension d'une boucle de cyclotide (une petite protéine cyclique de plante) par greffe de peptide responsable de l'antagonisme du VEGF, facteur de croissance de l'endothélium vasculaire (Gunasekera et al., 2008).

I.3.3.3. Transferts depuis des interacteurs naturels

Avec la disponibilité croissante d'informations expérimentales sur des interactions protéine-protéine naturelles, dans des contextes physiologiques ou pathologiques, l'identification de domaines fonctionnels naturels a été de plus en plus fréquente. De ce fait, les résidus fonctionnels de ces interacteurs ont pu être greffés sur de nouvelles charpentes, dans des

approches complétées par du design computationnel (que nous aborderons plus amplement dans la section "I.4. Approches computationnelles de design rationnel") pour générer une structure favorable à la présentation des motifs transférés.

C'est notamment le cas de travaux réalisés par Correia et ses collaborateurs ayant permis de transférer un motif discontinu de la surface de gp120 vers une charpente protéique découverte dans la PDB à l'aide de *Multigraft Match* (Azoitei et al., 2011). Malgré des affinités de plus de 300 μ M après le design initial, les auteurs ont pu atteindre des K_D jusqu'à la dizaine de nanomolaires après maturation d'affinité via des banques rationnelles, tout en fixant spécifiquement l'anticorps ciblé par l'épitope transféré. Le même groupe a également décrit la greffe de résidus fonctionnels sur des charpentes naturelles identifiées dans la PDB ou générées *de novo* par design computationnel, pour créer des vaccins à l'encontre d'épitoopes faiblement immunogènes du VIH (Correia et al., 2010) et du virus respiratoire syncytial (Correia et al., 2014), ou pour bâtir un inhibiteur du virus d'Epstein-Barr (Figure 10; Procko et al., 2014).

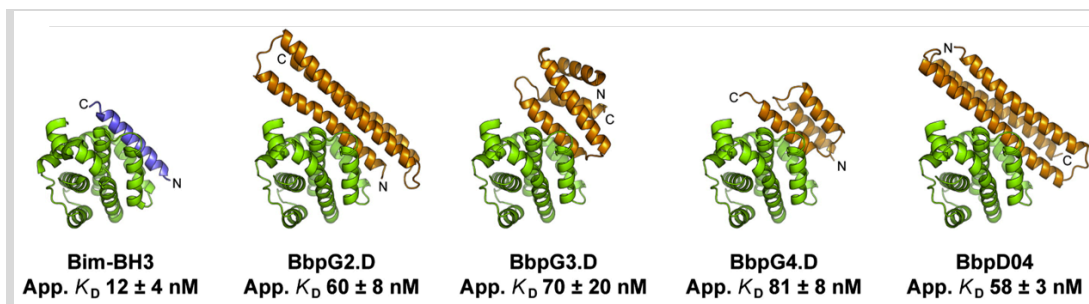
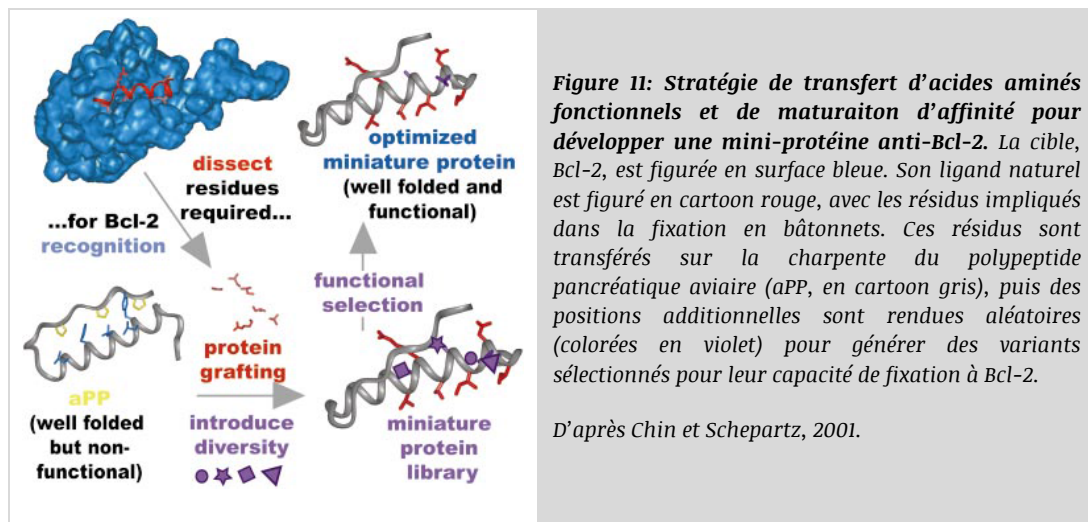


Figure 10: Modèles de charpentes en hélices (en orange) greffées de résidus fonctionnels d'un ligand naturel (en bleu) pour se lier à son récepteur (en vert). La structure avec la protéine Bim-BH3 correspond à la structure cristallographique avec son récepteur. Celles avec les protéines BbpG sont issues des modèles de transfert sur charpentes identifiées dans la PDB. Celles avec les protéines BbpD sont issues des modèles de transfert sur des protéines assemblées *de novo*. Les K_D apparent sont indiqués sous chaque structure, tels que déterminés par yeast display. Adapté d'après Procko et al., 2014.

Ces exemples récents indiquent que des transferts de résidus fonctionnels peuvent être réalisés en dehors de boucles, référence instaurée par l'humanisation des anticorps. Les hélices avaient d'ailleurs été privilégiées auparavant dans plusieurs études du transfert de résidus impliqués dans la formation d'interfaces protéine-protéine, en particulier dans le laboratoire de Schepartz et ses collaborateurs avec l'utilisation d'un polypeptide pancréatique aviaire comme charpente de mini-protéine (Chin et Schepartz, 2001; Rutledge et al., 2003). Dans ces premiers travaux, les auteurs ont généré une protéine spécifique de Bcl-2 ou d'un domaine de liaison à la protéine CREB

en greffant des résidus issus de leur ligand naturel sur la mini-protéine, puis ont introduit une diversité par randomisation de positions supplémentaires, pour finalement sélectionner et cribler les variants fonctionnels par *phage display* (Figure 11). Plus récemment, un domaine anti-albumine d'origine bactérienne a également été transféré sur six charpentes humaines en hélices pour leur procurer une extension de demi-vie sans nécessiter de fusion (stratégie préalablement abordée dans la section "1.2.3. Demi-vie") par greffe de 13 à 15 résidus (Oshiro et Honda, 2014). L'introduction de ces mutations n'a pas eu d'impact délétère sur la structure des protéines receveuses, et a permis d'obtenir des interacteurs effectivement compétiteurs du module anti-albumine non greffé (avec la meilleure affinité obtenue à 100 nM).



Cet attrait pour les greffes de résidus fonctionnels sur des hélices s'est également manifesté dans des transferts sur des hélices en glissière à leucine, motifs structuraux permettant la dimérisation de longues hélices alpha. Par exemple, cet agencement protéique a été l'hôte du transfert de l'épitope complet d'un peptide du VIH (situé à l'extrémité C-terminale de gp41), étendu sur 19 acides aminés non consécutifs (Sia et Kim, 2003), pour tenter de bloquer la voie d'entrée du virus dans les cellules. La protéine greffée a été décrite avec une efficacité proche de celle du ligand natif, en étant correctement structurée, stable sur le plan conformationnel et présentant une résistance accrue aux protéases, le tout sans optimisation après transfert. Les glissières à leucine ont aussi été décrites pour le transfert d'un nombre plus restreint de résidus, notamment lors de la génération d'inhibiteurs dimériques des protéines de la famille HDM après greffe de 3 résidus du domaine de transactivation de p53 (J.-H. Lee et al., 2014). La protéine résultante, p53LZ2, a

montré une fixation spécifique des cibles HDM proche de celle de p53 natif, confirmée par caractérisation structurale (**Figure 12**).

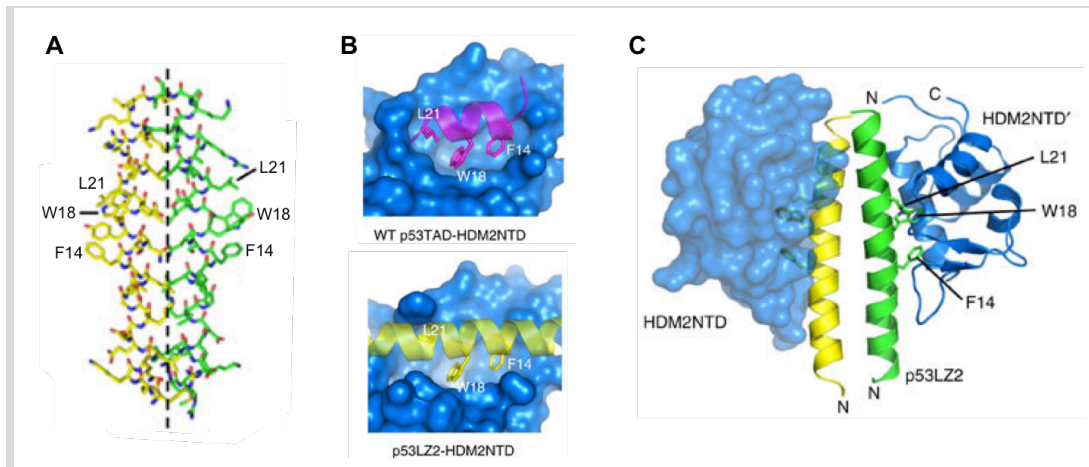


Figure 12: Homodimérisation et fixation d'un ligand de p53 après greffe de 3 résidus fonctionnels sur une glissière à leucine. A) Structure de l'homodimère p53LZ2 (chaînes jaune et verte), avec les 3 résidus hydrophobes F14, W18 et L21 indiqués. L'axe de dimérisation est représenté en pointillés. B) Comparaison des structures de p53 (magenta) et p53LZ2 (jaune) lié à la poche hydrophobe de HDM2 (surface bleue). Les résidus hydrophobes greffés de p53 à p53LZ2 sont indiqués. C) Structure du complexe de p53LZ2 homodimérique (chaînes jaune et verte) lié à deux molécules de HDM2 (en surface ou en cartoon bleu), avec les résidus hydrophobes responsables de l'interaction représentés en bâtonnets. Adapté d'après J.-H. Lee et al., 2014.

I.3.3.4. Transferts de feuillets beta

Quelques exemples de transferts à la surface de feuillets beta ont été décrits à ce jour par Guptasarma et ses collaborateurs, illustrant la capacité à créer des protéines hybrides dont les résidus enfouis des feuillets proviennent d'un homologue thermophile, tandis que les résidus exposés à la surface des feuillets assurent la fonction de la protéine (Kapoor et al., 2008, 2009). En résumé, ces greffes de résidus de surface trouvent leur succès dans deux critères structuraux essentiels. Tout d'abord, les squelettes peptidiques des protéines donneuse et receveuse doivent être superposables (avec un RMSD $\leq 2,0$ Å). Ensuite, les surfaces exposées à greffer doivent être majoritairement articulées autour de structures en feuillets beta, avec des boucles inter-brins s'organisant le moins possible en hélice. Grâce à cette approche, les auteurs ont réussi à mettre à profit la capacité des feuillets à conserver leur polarité entre faces enfouies et exposées malgré l'introduction de nombreuses mutations (**Figure 13**), comme nous avons également cherché à le réaliser lors de l'approche d'humanisation de Nanofitines décrite dans le chapitre "Chapitre II: Humanisation de Nanofitines par greffe de domaine".

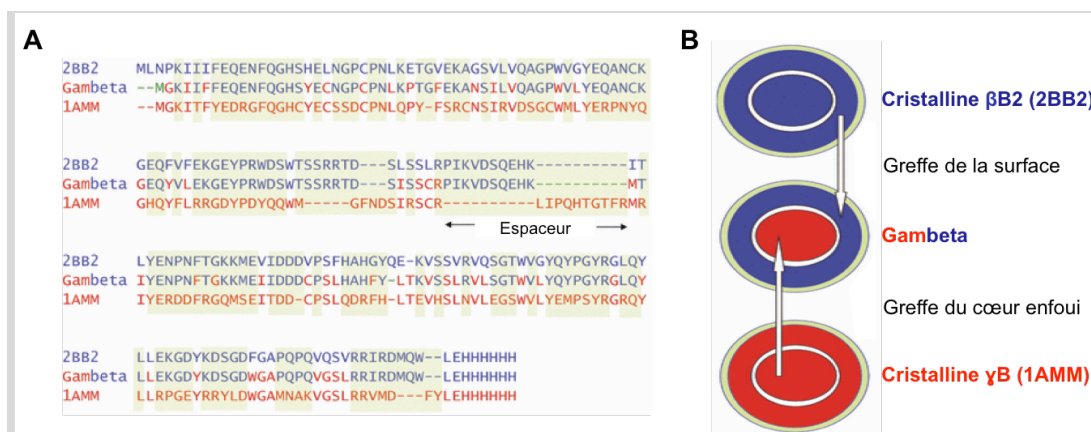


Figure 13: Construction de Gambeta, protéine composée de feuillets beta dont la surface accessible au solvant et le cœur enfoui ont été greffés, respectivement, depuis les cristallines β B2 (en bleu) et γ B (en rouge). A) Alignements de séquences. B) Représentation schématique du transfert de surface protéique. Adapté d'après Kapoor et al., 2009.

1.4. Approches computationnelles de design rationnel

1.4.1. Protéines de liaison et reconnaissance moléculaire spécifique

En supplément d'une approche d'humanisation de Nanofitines, Affilogic a souhaité augmenter la maîtrise du ciblage précis des épitopes ciblés par ses protéines de liaison générées par sélection *in vitro*. Les banques de Nanofitines employée lors de ce procédé allant jusqu'à 14 positions randomisées (encodant une diversité théorique de plus de $1,6 \times 10^{18}$ variants), elles peuvent dépasser la capacité maximale de criblage estimée à 10^{14} pour les techniques purement *in vitro* comme le *ribosome display* (Schaffitzel et al., 2001). Même si des banques réduites peuvent contourner cette limitation (dont la banque à 11 positions exploitée dans notre stratégie d'humanisation), l'exploration d'une telle diversité présente des biais intrinsèques dont ceux issus de la stabilité des protéines traduites, ou de l'accessibilité et de la facilité de ciblage d'épitopes distincts à la surface de la cible. Il n'est donc pas rare de pouvoir observer une convergence non désirée vers des épitopes privilégiés, notamment si la stratégie de sélection et de criblage ne comprend pas de contraintes suffisantes (Even-Desrumeaux et al., 2014). A l'instar des mécanismes d'évolution dirigée, nous pouvons représenter ce phénomène par des déplacements dans un paysage adaptatif, qui peuvent converger vers des maxima correspondant ou non à l'activité désirée (Figure 14).

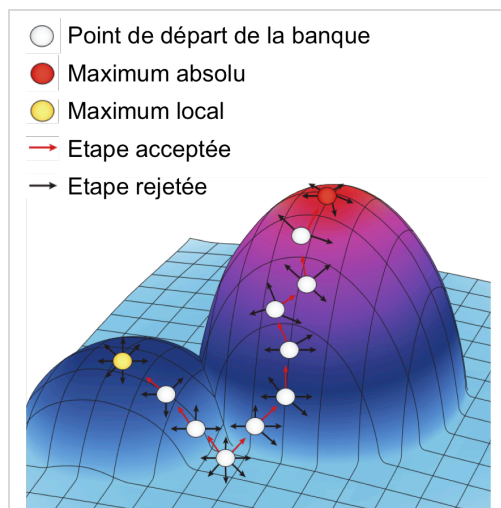


Figure 14: Paysage adaptatif exploré par les banques de variants. L'évolution dirigée peut être visualisée comme une série d'étapes dans un paysage adaptatif en trois dimensions. La génération de banque explore la surface proximale du paysage, et le criblage ou la sélection identifient des moyens "d'escalader" vers les pics d'adaptation. Le processus d'évolution dirigée peut aboutir à des niveaux d'activité maximale absolue, mais peut aussi rester piégé aux maxima locaux d'adaptation dans lesquels la diversification de la banque est insuffisante pour traverser des "vallées d'adaptation" et accéder aux pics d'adaptation voisins.

Adapté d'après Packer et Liu, 2015.

Il est souhaitable de pouvoir contrôler au mieux ce phénomène de convergence, en particulier pour la génération de protéines d'affinité à usage thérapeutique. En règle générale, la maîtrise de cette convergence lors du processus de sélection a pour effet de limiter l'effort de criblage et peut donner accès à des variants efficaces et sélectifs. Le fait de pouvoir prédire l'épitope ciblé par une molécule de liaison est également un moyen d'anticiper son mode d'action (**Figure 15**), ce qui est par ailleurs mis à profit lors de l'utilisation de protéines d'affinité dont l'épitope reconnu joue un rôle crucial dans leur activité (en plus de leur affinité) dans le développement d'agents neutralisants devant inhiber des interactions entre partenaires moléculaires. Un des avantages de ces approches thérapeutiques de blocage, qui peuvent notamment viser des interactions intracellulaires mais également récepteur-ligand ou enzyme-ligand, réside dans le fait que le mécanisme effecteur est directement assuré par l'action de fixation et ne nécessite donc pas d'autre fonction effectrice (qui pourrait être toxique pour les tissus sains). L'accès à de telles propriétés de ciblage spécifique et de neutralisation justifie le développement important de protéines d'affinités alternatives aux anticorps comme agents bloquants, notamment comme molécules anti-cancéreuses (**Figure 16**).

A ce jour, Affilogic a identifié des Nanofitines ciblant efficacement des protéines de signalisation telles que le facteur de nécrose tumorale alpha (TNF α ; Cinier et al., 2015), le facteur de croissance de l'endothélium vasculaire (VEGF), l'interleukine 17 (IL-17) et des récepteurs membranaires (Prel, 2015). En plus de maximiser les chances de découverte d'agents bloquants, le contrôle précis de la surface avec laquelle interagir peut autoriser le ciblage d'épitopes communs à diverses cibles.

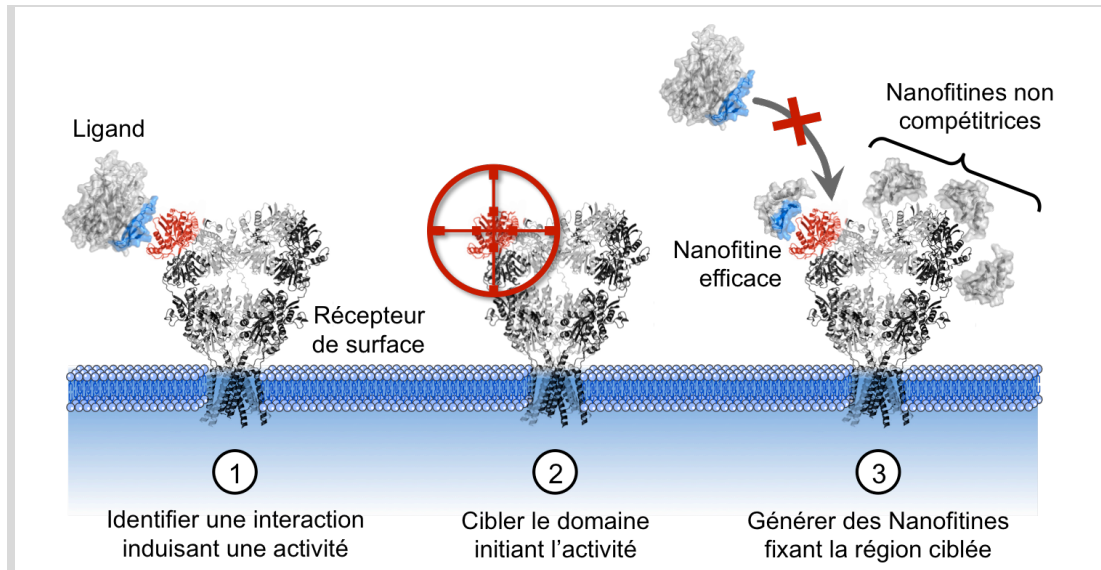


Figure 15: Prédiction de l'effet de Nanofitines comme agents bloquants en fonction de leur épitope. Exemple d'un récepteur de surface induisant une activité biologique lors de la fixation de son ligand naturel (1). Les surfaces d'interaction du récepteur et du ligand sont colorées, respectivement, en rouge et bleu. Le domaine d'interaction à la surface du récepteur peut être identifié à l'aide de données expérimentales ou prédictives, et être défini comme cible à atteindre (2). Les Nanofitines affines qui fixent le domaine ciblé sur le récepteur (avec la surface d'interaction représentée en bleu) sont attendues comme efficaces du fait de leur activité de compétition avec le ligand naturel, tandis que les Nanofitines se fixant sur d'autres régions du récepteur n'auront sans doute pas d'effet inhibiteur de l'interaction récepteur-ligand.

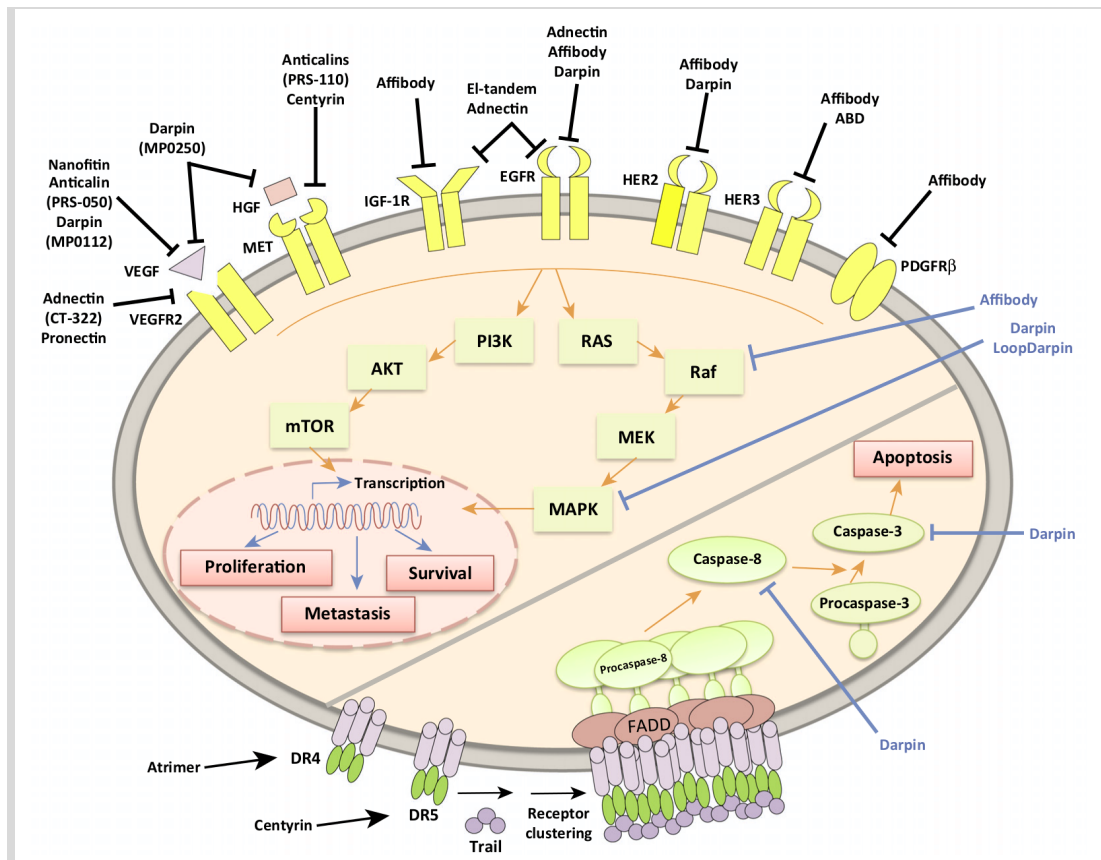


Figure 16: Agents bloquants non immunoglobulines actuellement décrits pour le traitement de cancers. Adapté d'après Škrlec et al., 2015.

C'est notamment une propriété désirée de molécules dirigées à l'encontre de pathogènes à fréquence de mutation élevée, pour le développement de diagnostics, vaccins ou traitements à large spectre. Un exemple marquant dans le domaine peut être illustré par le ciblage des virus grippaux, pour lesquels les régions les moins variables (assurant leur fonction) sont souvent moins accessibles, ce qui leur permet d'échapper aux défenses naturelles (Rossmann, 1989). L'identification de motifs viraux conservés et accessibles est à l'heure actuelle très active, permettant la génération régulière de molécules à large spectre spécifiques de l'hémagglutinine (Corti et al., 2011; Ekiert et al., 2009; Friesen et al., 2014; Mallajosyula et al., 2015; Sui et al., 2009). Nous soulignerons également le succès de la récente intégration de design protéique *de novo* pour générer une protéine anti-hémagglutinine d'affinité nanomolaire après maturation d'affinité *in vitro* (Fleishman et al., 2011), illustrant la capacité de prédiction que nous pouvons attendre de telles approches computationnelles.

Le ciblage de régions identiques ou homologues est aussi au cœur d'une problématique relative à la mise en place et à la réussite des tests pré-cliniques, étape obligatoire dans le développement de molécules thérapeutiques. En effet, la spécificité des molécules de fixation est un de leurs avantages majeurs (spécificité de séquence et d'arrangement tridimensionnel précis), mais présente le risque de ne pas pouvoir être applicable à plusieurs espèces. Cette limitation inhérente à l'utilisation de modèles animaux (revue par Pegram et Ngo, 2006) nécessite donc de pouvoir cibler efficacement des protéines orthologues pour pouvoir mener des études pré-cliniques fiables. Lorsque le criblage initial n'a pas permis de remplir cette condition, il est parfois possible de modifier les protéines d'affinité pour rétablir une réactivité croisée entre protéines humaines et du modèle animal, dont des exemples ont été illustrés par Werther et al., 1996; Liang et al., 2006; Garcia-Rodriguez et al., 2007. Encore une fois, il apparaît évident qu'une approche prédictive pourrait théoriquement répondre à ces attentes. Un exemple réussi d'une telle stratégie a été décrit en oncologie pour l'optimisation d'un anticorps dont une des huit mutations prédites a effectivement montré une augmentation d'affinité pour l'enzyme murine ciblée (multipliée par 14), aboutissant à une constante d'inhibition de 340 pM (Farady et al., 2009).

En opposition avec le ciblage d'épitopes communs, la reconnaissance spécifique d'épitopes distincts peut être primordiale pour une discrimination efficace de protéines homologues, notamment dans le cadre d'un diagnostic précis ou d'un traitement à spectre très étroit. Par exemple, pouvoir différencier un récepteur membranaire de sa contrepartie soluble (dont l'importance est soulignée dans la revue *Heaney et Golde, 1998*) peut présenter des avantages notables pour le traitement de pathologies. Actuellement, peu de cas sont documentés et se résument à des découvertes fortuites comme dans le cas d'anticorps ciblant le marqueur de différenciation CD4 soluble uniquement, permettant de mieux caractériser l'autoimmunité CD4 impliquée lors d'infections par le VIH (*Denisova et al., 2003*).

La capacité à discriminer des protéines homologues est d'autant plus importante dans le cas de la formation de multimères, comme c'est le cas de nombreux récepteurs ou cytokines. En considérant par exemple les membres de la famille de l'interleukine 12 (IL-12), la molécule agoniste active (IL-12p70) est un hétérodimère constitué des deux sous-unités IL-12p35 et IL-12p40 (*Gubler et al., 1991*). Son ciblage doit donc être suffisamment spécifique pour ne pas reconnaître son antagoniste formé par l'homodimérisation de la sous-unité IL-12p40 (*Gillessen et al., 1995*), ou bien l'interleukine 23 (IL-23) générée par l'hétérodimérisation de IL-12p40 et IL-13p19 (*Oppmann et al., 2000*).

Enfin, il est parfois bénéfique de pouvoir stabiliser des assemblages protéiques pour répondre à des problématiques de formulation de médicaments. Ce genre d'approche peut également mener à l'obtention de structures de complexes par cristallographie aux rayons X, via l'utilisation de petites protéines d'affinité comme chaperonnes (pour des revues de l'utilisation de fragments d'anticorps et charpentes alternatives en cristallographie, voir *Griffin et Lawson, 2011*; et *Sennhauser et Grütter, 2008*). En plus du criblage des protéines stabilisantes, comme la DARPin anti-AcrB se liant à la jonction de plusieurs sous-unités (**Figure 17**), il serait intéressant de pouvoir obtenir des protéines reconnaissant les interfaces des complexes de façon rationnelle.

Les exemples, non exhaustifs, que nous avons présentés ici montrent la finesse qui peut être nécessaire dans l'établissement d'une reconnaissance moléculaire entre protéines, aussi bien en

matière d'affinité que de spécificité, pour pouvoir tirer profit de protéines de liaison efficaces. Dans toutes ces illustrations, il apparaît également attrayant de bénéficier de méthodes prédictives performantes de design, dont nous allons faire état dans la suite de cette section.

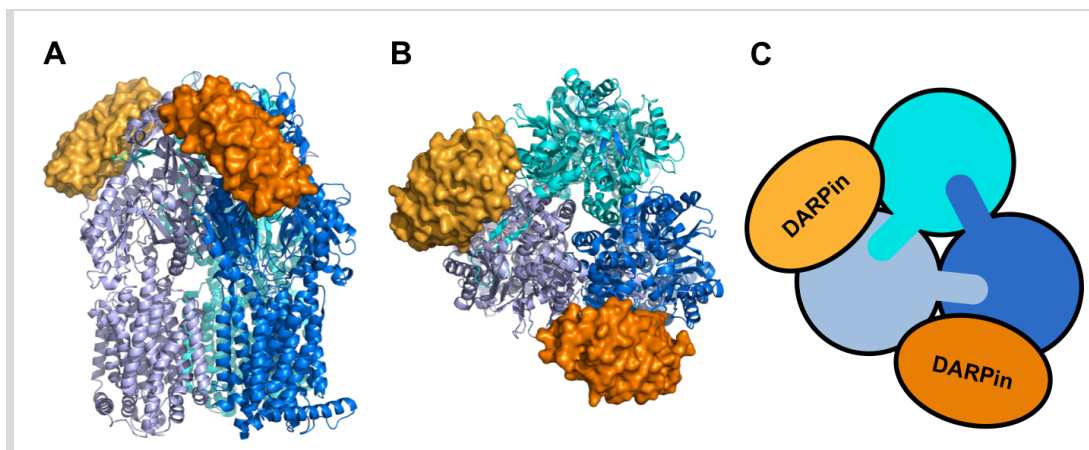


Figure 17: Exemple de stabilisation de complexe via une protéine d'affinité. Structure publiée sous l'identifiant PDB: 2J8S; Sennhauser et al., 2007. Les sous-unités de la protéine de *E. coli* AcrB sont représentées en dégradé de bleu, et les DARPins en dégradé d'orange. A) Vue latérale du complexe. B) Vue depuis le périplasma. C) Schématisation de la vue depuis le périplasma.

I.4.2. État de l'art en design protéique *in silico*

I.4.2.1. Généralités

D'un point de vue général, nous attendons des algorithmes de design *in silico* d'interfaces protéine-protéine qu'ils puissent mimer au mieux l'évolution dirigée des protéines (dont les méthodes sont rappelées de manière intéressante dans la revue Packer et Liu, 2015), tout en prenant en compte les contraintes spécifiques que nous imposons (épitope à cibler, orientation du complexe, composition de séquence, etc.). En fait, ces algorithmes doivent pouvoir simuler fidèlement l'évolution des interfaces protéine-protéine au sens large, puisqu'il a été suggéré que les expériences de maturation d'affinité sur des surfaces prédéfinies suivent les mêmes voies que l'évolution naturelle de ces interfaces (B. Li et al., 2010). Il a même été avancé plus récemment que les interfaces de type anticorps-antigène sont en réalité plus proches des contacts non corrélés fonctionnellement que des interfaces liées à la fonction des protéines, du fait de pressions de sélection différentes (Sudarshan et al., 2014).

Ces informations sont à traduire fidèlement sous forme de modèles (compaction, électrostatique, solvatation, van der Waals, rotamères, etc.) pour être calculées et pour diriger le design de

nouvelles protéines. En effet, la modélisation de la structure de protéines d'affinité (historiquement, les anticorps) a permis de conduire des approches rationalisées de maturation d'affinité au cours des 20 dernières années, guidées par la bioinformatique. C'est notamment le cas d'une des premières études dans le domaine (*Balint et Larrick, 1993*), estimant qu'une analyse de séquence et des informations de structures tridimensionnelles pouvait aider à la découverte d'anticorps de meilleures affinités par rapport aux approches d'immunisation d'animaux, en anticipant les possibilités de contrôler les réactivités croisées et faciliter l'humanisation des anticorps. Depuis, le nombre de méthodes de modélisation moléculaire s'est multiplié et celles-ci ont été appliquées dans différents domaines, dont le design d'enzymes et d'anticorps.

I.4.2.2. Méthodes disponibles et exemples

Les logiciels de design protéique actuellement disponibles sont variés, et une partie d'entre eux sont rendus directement accessibles par l'intermédiaire de serveurs en ligne. Parmi ces logiciels, nous pouvons citer dans une liste non-exhaustive: Rosetta (*Leaver-Fay et al., 2011*), ORBIT (*Dahiyat et Mayo, 1997*), FoldX (*Schymkowitz et al., 2005*), CC/PBSA (*Benedix et al., 2009*), IPRO (*Saraf et al., 2006*), OptCDR (*Pantazes et Maranas, 2010*) ou OptMAVEN (*T. Li et al., 2014*), OSPREY (*Gainza et al., 2013*), Proteus (*Simonson et al., 2013*), ZEMu (*Dourado et Flores, 2014*), AbDesign (*Lapidoth et al., 2015*), etc. Dans ce manuscrit, nous illustrerons majoritairement l'utilisation de la suite Rosetta, que nous avons sélectionnée du fait de ses succès récurrents en matière de prédiction de repliement et de design *de novo* (*Khoury et al., 2014*), ainsi que de modification de reconnaissance moléculaire entre protéines (*London et Ambroggio, 2013*; avec 2 à 20% de taux de succès pour le design d'interactions protéine-protéine).

Des illustrations intéressantes de la capacité de prédiction de structure et de fonction de protéines ont été obtenues dans le cadre de design *de novo* d'enzymes. Cette ingénierie rationnelle présente des défis de précision, car l'introduction de mutation doit diriger des modifications de squelette peptidique vers une conformation idéale pour présenter les chaînes latérales fonctionnelles dans les bonnes orientations. En plus de revues récentes sur les outils computationnels pour le design et l'ingénierie d'enzymes (*Zanghellini, 2014*; *Damborsky et*

Brezovsky, 2014), nous pouvons souligner l'excellent taux de succès observé lors du design *de novo* d'enzymes rétro-aldol, ayant abouti à 32 mutants actifs sur les 72 testés à partir de charpentes protéiques diverses, dont deux validées après cristallographie aux rayons X (Jiang *et al.*, 2008). Un élément clé du succès de ces prédictions, selon les auteurs, réside notamment dans la modélisation explicite des molécules d'eau au sein du site actif, souvent primordiales et conservées (Teze *et al.*, 2013).

La même année a vu la parution du design *de novo* d'enzymes catalysant avec succès une réaction d'élimination de Kemp, une réaction modèle de transfert de protons à partir de carbones non catalysée par des enzymes naturelles (Röthlisberger *et al.*, 2008). Les enzymes prédites, en plus d'avoir été confirmées par cristallographie, ont donné accès à une vitesse d'accélération de la réaction $k_{\text{cat}}/k_{\text{uncat}}$ de 10^5 avec une constante de spécificité k_{cat}/K_M de 100 (encore multipliée par 200 par évolution dirigée *in vitro*). Une approche similaire a permis au même groupe de générer des enzymes catalysant des réactions de Diels-Alder, formant deux liaisons carbone-carbone en une seule étape, non réalisées par des enzymes naturelles (Siegel *et al.*, 2010).

Un autre exemple de contrôle de la spécificité d'une enzyme a été réalisé par réagencement et design d'une boucle autour du site actif (Murphy *et al.*, 2009). Après avoir validé l'approche sur huit enzymes références, les auteurs l'ont appliquée au redesign d'une guanine désaminase humaine et ont réussi à augmenter l'activité à partir d'un substrat ammélide par 100 tout en réduisant celle à partir de la guanine par un facteur $2,5 \times 10^4$, multipliant ainsi la spécificité du substrat non naturel de l'enzyme par un facteur $2,5 \times 10^6$ par rapport à son substrat naturel. Leurs prédictions ont par ailleurs été confirmées par l'obtention d'une valeur de RMSD inférieur à 1 Å en comparaison avec la structure cristallographique.

Le design de boucles est également très répandu pour l'optimisation assistée par modélisation des régions déterminant la complémentarité des paratopes d'anticorps, constituées de 3 boucles variables par chaîne, comme résumé dans la revue Kuroda *et al.*, 2012. Le défi de l'exploration conformationnelle de ces boucles est d'autant plus important pour la troisième CDR des chaînes lourdes (CDR-H3), car cette dernière est plus longue et peut donc adopter des arrangements plus

diversifiés (**Figure 18**). Entre autres, Gray et ses collaborateurs ont proposé l'implémentation de protocoles dans Rosetta pour modéliser les boucles CDR-H3 par assemblage de fragments, descente cyclique de coordonnées et minimisation (Sivasubramanian *et al.*, 2009), avec des prédictions suffisamment fiables pour réaliser des simulations d'ancrage moléculaire et de design d'interface par la suite (comme nous l'aborderons dans la section "I.4.3. La suite de modélisation Rosetta"). En plus des groupes de Gray ou Kortemme (entretenant le développement de Rosetta), des groupes comme celui de Maranas et ses collaborateurs développent des alternatives telles que MAPs, une banque structurale d'anticorps pour la prédiction de structure 3D (avec 1,9 Å de RMSD en moyenne sur tous les atomes de 260 anticorps testés) et la maturation d'affinité consécutive, intégrée à la plateforme IPRO mise à disposition des chercheurs (Pantazes *et al.*, 2015; Pantazes et Maranas, 2013).

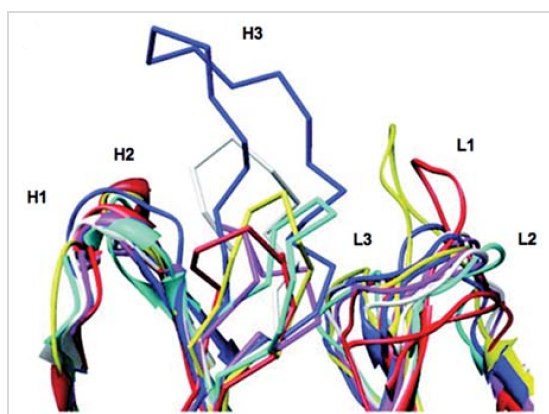


Figure 18: Diversité structurale du paratope des anticorps, illustrée par superposition de 5 structures. Les CDR sont nommées de H1 à H3 et de L1 à L3, respectivement, pour les chaînes lourdes et légères. Chaque immunoglobuline est représentée d'une couleur différente.

D'après Kuroda *et al.*, 2012.

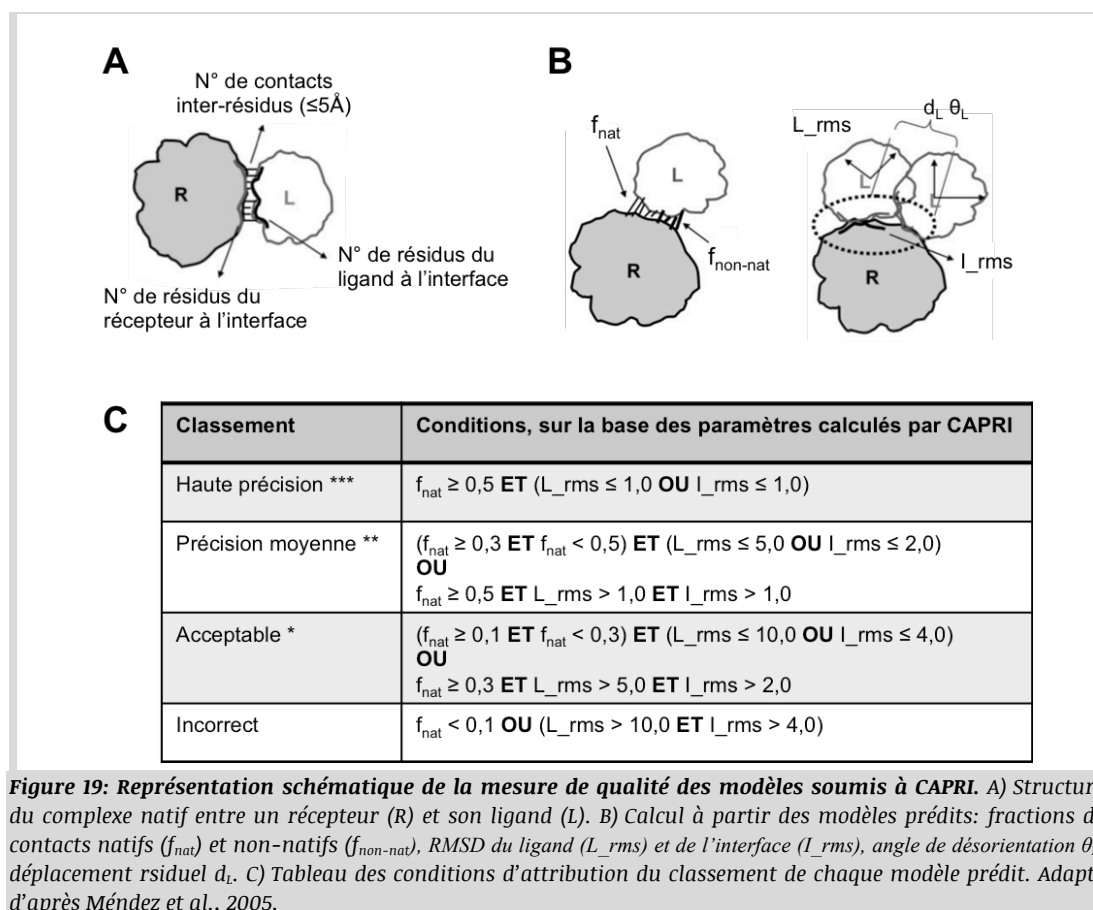
A proprement parler, nous constatons tout de même que la majorité des approches actuelles de design de boucles pour contrôler une reconnaissance moléculaire anticorps-antigène, en particulier pour effectuer des maturations d'affinité *in silico*, constituent du re-design d'interface puisqu'elles démarrent de structures de complexes disponibles dans la PDB. Ainsi, les affinités de diverses interfaces ont pu être optimisées par des approches computationnelles, dont celle d'un anticorps dirigé contre le domaine-I de l'intégrine VLA1, initialement estimée à 7 nM puis jusqu'à 850 pM après optimisation (sur plus de 80 variants testés; Clark *et al.*, 2006). Cette étude a confirmé les contacts prédits et souligné l'importance de l'établissement de liaisons hydrogène pour la recherche d'affinités accrues.

I.4.2.3. L'expérience CAPRI

Pour réaliser du design *de novo* au sens strict, un challenge additionnel est de déterminer un point de départ à l'interface à optimiser. L'importance de la définition de ce point de départ, aussi bien en matière de surfaces mises en contact que d'orientation des complexes, implique l'exploration d'un large espace conformationnel et souligne le rôle crucial des simulations d'ancrage moléculaire. Ce constat figure dans les motivations du test CAPRI (*Critical Assessment of Predicted Interactions*), auquel tous les participants volontaires sont invités à participer (Janin, 2002). En résumé, le concept de CAPRI est de tester les méthodes et le savoir-faire de la communauté à l'égard des prédictions de complexes protéine-protéine pour identifier la structure des complexes et leurs interfaces avec une précision atomique. Lors de chaque édition (ou tour) de CAPRI, les organisateurs mettent à disposition des participants des structures de complexes récemment résolus, mais pas encore publiés (sur la base de la contribution d'expérimentateurs). Ensuite, les participants bénéficient de quelques semaines pour reconstituer les complexes à partir des structures des monomères non liés (voir de monomères extraits de la structure d'un complexe, dans le cas où un des monomères non liés n'est pas disponible) en utilisant des informations structurales, des critères d'énergie, de conservation de séquence ou des combinaisons variées de critères et d'approches. Les modèles proposés sont finalement évalués et jugés, tel que résumé sur la **Figure 19**, comme “*** haute précision”, “** précision moyenne”, “* acceptable” ou “incorrect”, sur la base de trois mesures: la fraction des contacts natifs inter-résidus (f_{nat}), le RMSD de la position des carbones alpha du partenaire le plus petit après superposition de la protéine de plus grande taille (L_{rms}), et celui des carbones alpha de l'interface (I_{rms}). L'approche proposée par CAPRI est donc avantageuse pour tester la robustesse d'une approche de prédiction de complexe, puisqu'elle permet la validation directe des résultats par les organisateurs, au lieu de la laisser à la charge des auteurs des prédictions.

Les responsables des évaluations des prédictions proposées dans le cadre du projet CAPRI, Wodak et Lensink, ont donc régulièrement l'opportunité de faire un bilan comparatif des groupes et méthodes participants, et faire ainsi un tour d'horizon des challenges relevés et de ceux à surmonter. Au tour 19 (correspondant à l'évaluation des 42 premières cibles proposées par CAPRI;

Lensink et Wodak, 2010), ils ont pu évaluer les performances limitées des différentes approches lorsqu'il s'agit de prédire des structures de complexes, en particulier ceux dits "facultatifs" correspondant à l'interaction entre deux partenaires stables individuellement. Cependant, ils ont également pu mettre en évidence une tendance générale à identifier correctement les interfaces ciblées (avec une précision d'environ 60%, et des variabilités notables entre les participants et les cibles considérés). Dans le cadre de l'étude de design *de novo* que nous voulons mettre en place autour des Nanofitines, nous nous sommes particulièrement focalisés sur la suite Rosetta, ayant fait preuve de succès récurrents lors des 30 premiers tours de CAPRI par l'intermédiaire des prédicteurs humains et des serveurs de Rosetta en ligne dirigés, respectivement, par les groupes de Baker et Gray.



Lors des tours 3 à 5 de CAPRI (Daily et al., 2005), les premiers tests d'ancrage moléculaire obtenus avec Rosetta ont permis de prédire 3 complexes avec une haute précision, 2 avec une précision moyenne et 1 complexe acceptable, à partir de 9 complexes (dont une majorité possédait au

moins un des partenaires avec des informations biologiques disponibles). Les auteurs ont principalement mis en avant leur difficulté face à des protéines de plus grande taille (≥ 450 résidus), car celles-ci nécessitaient un échantillonnage plus intense de l'espace à explorer. En réponse à ces observations, ils ont intégré une approche en deux temps dans laquelle les cycles de randomisation de type Monte-Carlo ont tout d'abord été effectués en approche gros-grain, c'est-à-dire à plus faible résolution en remplaçant les chaînes latérales des protéines par des représentations en centroïdes. Dans un second temps, ces cycles ont été effectués à plus haute résolution en considérant l'ensemble des atomes des partenaires, permettant d'affiner les simulations en ajoutant l'exploration des rotamères à celle des orientations globales des protéines.

Lors des tours 4 et 5 de CAPRI, Rosetta a permis de proposer la meilleure soumission pour 5 des cibles à prédire en couvrant l'ensemble de la surface accessible de la cible puis en affinant l'ancrage moléculaire localement (*Schueler-Furman et al., 2005*). Les prédictions proposées avec Rosetta ont généralement abouti à des résultats proches des complexes résolus par cristallographie aux rayons X, avec 8 cibles dont le RMSD du squelette peptidique à l'interface indique des prédictions de haute résolution (3 à moins de 0,4 Å et 6 à moins de 1,1 Å). De manière générale, les auteurs ont pu constater des succès en absence de mouvements significatifs du squelette peptidique des protéines étudiées (via une prédiction fidèle des modifications conformationnelles des chaînes latérales), et ont déploré les difficultés de leur approche pour obtenir de tels résultats lorsqu'une flexibilité importante était nécessaire pour assurer la transition entre les états libres et liés.

De ce fait, la suite Rosetta a été modifiée pour tenir compte de ces faiblesses et permettre de modéliser plus efficacement la flexibilité des protéines lors des tours 13 à 19 de CAPRI (*Fleishman et al., 2010*). En plus de l'évaluation des complémentarités géométriques des complexes à squelette peptidique fixe et de l'exploration des rotamères possibles, Rosetta a intégré des algorithmes supplémentaires permettant de prédire les mouvements des chaînes principales des protéines lors de la fixation. Ces changements conformationnels ont été appliqués avec des

amplitudes variables, du simple ajustement de boucles pour limiter les conflits stériques à la prédiction de structure flexible d'ARN. Grâce au succès avec un complexe trypsine-inhibiteur (meilleure pose soumise) et un complexe protéine-ARN, les auteurs ont pu confirmer leur capacité à identifier des complexes de haute et moyenne précision en simulant des mouvements des chaînes principales. Cependant, ils ont également identifié des problèmes d'échantillonnage aux extrémités exposées des brins beta et un point d'amélioration crucial (et en constante évolution): la nécessité de perfectionner les fonctions d'énergie.

I.4.3. La suite de modélisation Rosetta

I.4.3.1. Forces et améliorations apportées aux modèles prédits

L'expérience CAPRI montre Rosetta comme une suite logicielle pouvant donner accès à des prédictions de haute précision. Nous avons aussi précédemment abordé sa capacité à générer des protéines fonctionnelles par re-design et design *de novo*, indiquant la robustesse de cet outil pour réaliser des simulations d'ancrage moléculaire, de relaxation et de design. Notre approche de design *de novo* de Nanofitines nécessite typiquement les mêmes champs de force et les mêmes méthodes d'échantillonnage que pour le re-design, tout en nécessitant de pré-orienter les partenaires moléculaires en un complexe compatible avec la formation d'une interface spécifique et énergétiquement favorable. Dans cette section, nous justifierons donc ce choix en illustrant la polyvalence des applications de Rosetta (actuellement en version 3.x; *Leaver-Fay et al., 2011*), en sélectionnant quelques-unes des ~380 publications actuellement révisées impliquant cette suite de modélisation (seule, ou en combinaison avec d'autres techniques comme l'algorithme ZDOCK, outils concurrent mais aussi complémentaire; *Vreven et al., 2013*).

Comme nous l'avons évoqué préalablement, les techniques *in silico* semblent souffrir de limitations liées aux dimensions de l'espace conformationnel à explorer ainsi qu'aux imprécisions inhérentes aux fonctions d'énergie employées. Ceci peut d'ailleurs expliquer pourquoi les premières applications de design d'interface protéine-protéine avec Rosetta ont été centrées sur la prédiction de l'effet de mutations en alanine. En 2002, Kortemme et Baker ont montré la mise en place d'un algorithme prédisant avec succès la contribution de différents

acides aminés à la stabilité de protéines globulaires et de 19 interfaces entre protéines (743 et 233 mutations au total) avec respectivement des erreurs en valeur absolue de 0,81 et 1,06 kcal/mol, en comparaison avec les données expérimentales des mutations en alanine (Kortemme et Baker, 2002). Au niveau des interfaces protéine-protéine, ces résultats ont permis d'identifier 79% des points chauds et 68% des résidus neutres, ce qui a motivé les auteurs à mettre à disposition leur serveur en ligne de "cartographie par mutations computationnelles en alanine" (Kortemme, Kim, et al., 2004). L'amélioration et la comparaison des fonctions de score est également régulièrement étudiée par les groupes de modélisation moléculaire, avec notamment la proposition d'outils de comparaison de scores entre prédictions et structures résolues par cristallographie aux rayons X (Leaver-Fay et al., 2013).

I.4.3.2. Résolution de structure

Historiquement, la vocation initiale du logiciel de modélisation Rosetta a été décrite pour la prédiction de structures de protéines *ab initio*, c'est à dire la modélisation de la structure tertiaire de protéines à partir de leur séquence en acides aminés. Ainsi, les premiers succès ont été générés à partir de petites protéines en hélices, considérées comme un premier exemple de modèles correctement prédits du fait d'une structure ternaire globale proche des structures cristallographiques, malgré des RMSD de 7,2 Å en moyenne (Simons et al., 1997). Dans cette même étude, les structures organisées en feuillets ou boucles ont semblé bien plus difficiles à prédire. La méthode a été ensuite testée dans le cadre du projet d'évaluation des prédictions de structures CASP 3, à partir de 21 petites protéines (Simons et al., 1999), avec des RMSD entre 3,8 et 6,4 Å par rapport aux structures natives, pour des domaines protéiques de 67 à 99 résidus. Cinq éditions plus tard (CASP 8), Rosetta a permis de proposer de meilleurs modèles que les homologues présents dans la PDB pour 24 des 50 domaines à étudier. L'efficacité de Rosetta est surtout à souligner lors de l'utilisation des modèles de raffinement en prenant en compte tous les atomes: certaines des structures prédites *de novo* étaient meilleures que celles obtenues par homologie, et Rosetta a permis d'améliorer 7 des 12 modèles de structure à raffiner (Raman et al., 2009).

En dehors des modèles de CASP, Rosetta a été utilisée pour s'adresser à un problème structural majeur: la résolution de structure de protéines membranaires. Malgré de récentes avancées en la matière (Kang et al., 2013), la cristallographie de ces protéines présente de nombreuses contraintes (taille importante des protéines, expression et purification homogène délicates dans leur environnement lipidique, risque d'instabilité, etc.). Les méthodes de prédiction de structure de Rosetta pourraient compléter les données expérimentales acquises (structures d'homologues, effets de mutations, etc.) et éclairer la compréhension de mécanismes liés à ces protéines. Par exemple, la prédiction *de novo* a été effectuée sur des protéines transmembranaires hélicoïdales, avec des résultats intéressants en simulant un repliement séquentiel des hélices, et non pas en un seul bloc comme réalisé avec les protéines solubles (Yarov-Yarovoy et al., 2006). Sur les 12 protéines testées, les auteurs ont prédit entre 51 et 145 résidus à moins de 4 Å de RMSD par rapport à la structure native, identifiant par la même occasion un verrou technologique que nous avons décrit précédemment: la notion de flexibilité du squelette peptidique. Dans cette étude, la difficulté accrue à modéliser cette flexibilité a montré les limites des algorithmes de Rosetta à cette époque, notamment lorsque le nombre d'hélices ou le nombre et la longueur des boucles qui les séparent augmente. Un exemple plus récent concerne la prédiction de structure du canal potassique hERG1 complet (pore et domaine tensiosenseur; Subbotina et al., 2010). Face à la difficulté de cristalliser ce canal, les auteurs ont mis en place une méthode de combinaison entre prédiction de structure par homologie, prédiction *de novo* des domaines non cristallisés, et dynamique moléculaire en couplage avec les données expérimentales. Sur cette base, ils ont pu proposer des pistes quant aux mécanismes d'action du récepteur, et suggérer la présence de sites distincts pour le développement d'agonistes ou d'inhibiteurs du canal. Cependant, cette étude n'a pas été confirmée par la résolution de la structure de la protéine, et demande donc validation.

D'autres prédictions de structures jugées exotiques ont été menées via Rosetta, notamment avec des protéines liées à des métaux. En 2010, une prédiction des structures de protéines de liaison au zinc a notamment été effectuée sur différentes charpentes, exploitant des modèles explicites de la géométrie du polyèdre de coordination du métal (C. Wang et al., 2010). Grâce à cela, les auteurs ont mis en évidence des structures prédites avec des RMSD par rapport aux structures natives

allant de 2,89 à 11,99 Å en absence de zinc, et diminuant entre 0,70 et 7,35 Å grâce à la réorganisation induite en présence du métal capturé. Ces méthodes sont anticipées comme applicables à d'autres interactions avec des ions métalliques et/ou des co-facteurs chimiques aux géométries bien définies.

Dans tous ces applications (challenge méthodologique CASP, protéines solubles, protéines membranaires, protéines de coordination de métaux), les comparaisons cycliques entre les méthodes incrémentées dans la suite Rosetta et les résultats confirmés expérimentalement ont permis de définir les limitations des algorithmes utilisés pour chercher à les résoudre. Une des améliorations qui résultent de ce cheminement est le développement de RosettaHoles, une méthode d'évaluation de la bonne compaction des protéines (Sheffler et Baker, 2009, 2010). En identifiant des cavités (à l'aide de sphères de diamètres variables), les auteurs suggèrent la possibilité de discriminer les régions mal conformées dans les résultats de prédiction, mais également les régions possédant des résolutions locales trop faibles au sein de structures disponibles dans la PDB. La méthodologie a d'ailleurs été complétée par un groupe de chercheurs indépendants de Rosetta, pour implémenter une méthode rapide d'estimation des aires des surfaces accessibles au solvant (Durham et al., 2009).

A terme, l'amélioration de ces algorithmes devrait pouvoir augmenter la fiabilité des résultats de prédiction de structures, mais également permettre de raffiner des données expérimentales incomplètes ou de moindre résolution. C'est notamment le cas dans une publication de 2009, où la combinaison des algorithmes de repliement et d'ancrage moléculaire de Rosetta ont permis d'établir des modèles d'homo-oligomères proches des structures natives dans plus d'un tiers des cas étudiés (Das et al., 2009). Les auteurs suggèrent que ce type d'approche aboutissant à des résolutions proches de l'atome pourraient générer des modèles suffisamment fiables pour pouvoir s'en servir lors du phasage nécessaire à la résolution de structures cristallographiques (Wlodawer et al., 2013), ou pour déterminer une structure à partir d'informations limitées obtenues par résonance magnétique nucléaire. Ces méthodes proposées pour le raffinement de structures ont également été jugées appropriées pour générer des structures de haute résolution à

partir de cartes de densité de cryomicroscopie électronique (ayant de plus faibles résolutions que celles obtenues en cristallographie aux rayons X; DiMaio *et al.*, 2009).

A plus long terme, l'évolution de la précision des prédictions de Rosetta et des méthodes complémentaires existantes (comme nous pourrons le suivre à travers les futures éditions de CASP) laisse présager la possibilité de générer des modèles suffisamment fiables pour pouvoir s'y baser lors de démarches de design *de novo*. Toutefois, les limitations actuelles révélées par les évaluations de CASP et CAPRI tracent le chemin des améliorations restantes, mais il reste encore à le parcourir.

I.4.3.3. Ancrage moléculaire

Nous avons abordé l'évolution des compétences et des algorithmes de Rosetta pour l'ancrage moléculaire dans la section "I.4.2.3. L'expérience CAPRI". La démarche initialement adoptée (rappelée dans le cadre de CAPRI; Gray *et al.*, 2003), a consisté à simuler $\sim 10^6$ poses par complexe, en explorant les modifications conformationnelles des chaînes latérales et en opérant une minimisation avec squelette peptidique fixe. Les fonctions d'énergie employées étaient régies par des critères d'interactions de van der Waals, de modèle implicite de solvatation, et de potentiel de liaison hydrogène dépendant de leur orientation. Grâce à ces fonctions, considérées proportionnelles aux affinités expérimentales, les développeurs de Rosetta ont proposé une démarche simple de sélection des meilleures poses par tri selon leurs scores d'interaction, regroupement spatiaux, puis inspection manuelle. Pour cette approche d'ancrage, tout comme l'ensemble des prédictions générées par Rosetta, les auteurs ont également souligné la forte influence du facteur humain (possible dans l'analyse manuelle ou le choix des paramètres), pouvant alternativement améliorer ou fausser les résultats obtenus. Malgré tout, l'évolution des résultats de CAPRI suggère que la méthode est en bonne voie pour obtenir des modèles précis de complexes à partir de structures indépendantes, comme nous l'évoquerons ci-après à travers une série d'exemples mettant à profit ces démarches d'exploration conformationnelle (avec une attention particulière portée sur les défis méthodologiques qu'ils soulèvent).

Une des limitations récurrentes des algorithmes de prédiction structurale de protéines, y compris pour les simulations d'ancrage moléculaire, est la modélisation des liaisons électrostatiques et

en particulier l'établissement de liaisons hydrogène. Dès 2003, une des premières améliorations de la prédiction de ces liaisons hydrogène a été proposée pour la prédiction de complexes, en remplaçant le modèle de Coulomb par une fonction tenant compte des orientations de ces interactions, sur la base des observations réalisées sur les structures cristallographiques de haute résolution (Kortemme et al., 2003). Les auteurs ont confirmé l'intérêt d'une telle optimisation en réussissant à prédire de la sorte l'orientation de 22 complexes sur 31 testés (en utilisant trois termes pour les liaisons hydrogène), alors que seulement 13 succès étaient obtenus avec le modèle de Coulomb seul.

Un autre challenge majeur de la modélisation protéique concerne la prédiction de la flexibilité des protéines, en tenant compte de la relaxation possible du squelette peptidique. Après 2005, la méthode d'ancrage de Rosetta a été complétée par des protocoles de mouvements de chaîne principale (C. Wang et al., 2007), améliorant l'obtention de résultats proches des structures natives lorsque de légers mouvements sont nécessaires pour la formation de complexes. Pour tenir compte de mouvements plus importants, un algorithme de prédiction des conformations de boucles a été mis en place. En pratique, ces cas de larges réarrangements correspondent quasiment à deux prédictions de repliement *ab initio* combinées à une simulation d'ancrage moléculaire lors de l'utilisation de modèles flexibles. De ce fait, une attention particulière doit être portée lors du traitement des scores de ces méthodes, pour ne pas favoriser de complexes physiquement improbables (dépliage partiel d'un partenaire, interface réalisée avec des boucles entrecroisées, etc.). Par exemple, le score d'interaction ne doit pas être pris en compte seul lors de larges mouvements prédits par Rosetta, car ceux-ci peuvent modifier fortement le score total (reflétant la stabilité des partenaires). Malgré cette complexification de la méthode, cette implémentation présente un intérêt indéniable pour la prédiction de complexes protéine-protéine.

Pour illustrer concrètement l'usage de l'ancrage moléculaire comme outil de recherche, citons l'exemple de la prédiction du complexe entre un anticorps monoclonal (mAb 806, n'ayant pas de structure résolue) et le récepteur du facteur de croissance épidermique (EGFR) (Sivasubramanian

et al., 2006). Dans cette étude, les auteurs ont généré une structure par homologie et sur la base de mutagenèses expérimentales ou computationnelles pour proposer le mécanisme d'action de cet anticorps anti-tumoral, qui inhiberait la dimérisation du récepteur en se fixant à proximité de l'interface du dimère. Cette stratégie de prédiction de structure par homologie et d'ancrage moléculaire a été par la suite approfondie pour générer un algorithme baptisé *SnugDock*, spécialisé dans la prédiction de complexes anticorps-antigène (*Sircar et Gray*, 2010).

Les algorithmes d'ancrage que nous avons décrits ont été essentiellement développés et appliqués en direction de complexes macromoléculaires. Leur application semble possible pour de petites molécules et des peptides, comme en témoignent un protocole d'ancrage de petites molécules proposé en 2013 (*Combs et al.*, 2013) et une étude focalisée sur les interactions entre une protéine de type Bcl-2 et un domaine BH3 (*London et al.*, 2012). Dans ces travaux, les auteurs ont notamment démontré qu'un protocole d'ancrage de ligands à la surface d'une protéine permet de tester leur spécificité. Toutefois, d'autres méthodes existantes d'ancrage et de dynamique moléculaire (que nous ne décrirons pas dans ce manuscrit) semblent actuellement plus adaptées à ces problématiques engageant un partenaire de petite taille.

Pour finir, nous noterons que l'intérêt de l'ancrage moléculaire ne se limite pas à définir des surfaces d'interaction et des organisations géométriques de complexes uniquement pour leur observation et la compréhension de leur mécanisme. En effet, ces simulations peuvent être étroitement liées au design ou au re-design d'interfaces. D'une part, elles fournissent des poses de départ pour effectuer le design *in silico*. D'autre part, ce sont des fonctions d'énergie similaires qui sont impliquées en ancrage moléculaire ou en design (régulièrement affinées pour modéliser plus fidèlement les lois physiques ou optimiser les temps de calcul nécessaires).

I.4.3.4. Re-design

Le re-design tel que considéré dans ce manuscrit correspond à la prédiction computationnelle de l'impact de mutations sur des structures de protéines (et leur activité), à partir de structures initiales préalablement définies. Par exemple, des stratégies de re-design ont été opérées avec Rosetta pour augmenter la stabilité et la capacité de repliement de protéines (comme nous

l'avons abordé précédemment dans la section "1.2.2. Stabilité et solubilité" avec l'obtention de fragments scFv superchargés avec thermostabilité accrue; Miklos *et al.*, 2012). Ce type d'approche a été répété à nouveau dans l'optique de limiter l'agrégation et faciliter le bon repliement des protéines, notamment avec la GFP (Der *et al.*, 2013). Dans cette étude récente, les auteurs ont mis en avant une zone intermédiaire de charge nette permettant d'augmenter la stabilité de la GFP, tandis que des variants trop ou pas assez chargés ne bénéficiaient pas de cette amélioration. La méthode est d'ailleurs directement disponible en ligne sur le serveur ROSIE (Lyskov *et al.*, 2013), via l'outil *Supercharge*.

Pour l'application du re-design à l'ingénierie de protéines d'affinité, comme c'est l'objectif de notre étude autour des Nanofitines, il s'agit d'exploiter les méthodes disponibles par l'intermédiaire de RosettaDesign (également disponible sur un serveur en ligne, Yi Liu *et Kuhlman*, 2006), afin de modifier et stabiliser une interface protéine-protéine (et non plus seulement stabiliser un état de repliement). Plusieurs succès de re-design ont été réalisés à partir d'une DNase (la colicine E7) et son inhibiteur (Im7) (Kortemme, Joachimiak, *et al.*, 2004). Après avoir déstabilisé l'interface de 690 Å² (d'une affinité estimée entre 10⁻¹⁴ et 10⁻¹⁶ M) via l'introduction de mutations prédites, les auteurs ont volontairement obtenu des variants abrogeant l'interaction. Ensuite, un re-design a été effectué des deux côtés de l'interface (établie sur environ 750 Å² après mutations) pour identifier de nouveaux couples de variants spécifiques. Les mutants générés ont permis d'atteindre des affinités jusqu'à 0,34 nM. Sur le même modèle, un re-design intégrant des étapes d'ancrage moléculaire local a été réalisé avec un accent porté sur l'établissement de liaisons hydrogène (avant et après obtention de la structure cristallographique) pour permettre une forte augmentation de spécificité de l'interaction (×300 par rapport au complexe parental; Joachimiak *et al.*, 2006). En étudiant également ce complexe modèle dans des travaux ultérieurs (Sammond *et al.*, 2010), un autre groupe a approfondi cette étude via Rosetta, montrant à cette occasion la difficulté de basculer depuis des interactions hydrophobes vers des interactions électrostatiques, alors que l'inverse semble donner de meilleurs résultats.

Le même groupe a cherché à augmenter l'affinité d'interface dans des complexes pré-orientés, en identifiant des mutations ponctuelles (Sammond et al., 2007). Leurs prédictions structurales ont visé à ne pas déstabiliser les monomères, tout en augmentant la surface occupée par des résidus hydrophobes enfouis à l'interface et en maximisant le nombre de liaisons hydrogène pour les résidus polaires enfouis. Sur 12 mutations testées expérimentalement à partir de protéines ciblant leur partenaire non muté, les auteurs ont reporté 9 succès avec des affinités améliorées (jusqu'à $\times 26$), dont 5 présentant un gain de plus de 1 kcal/mol.

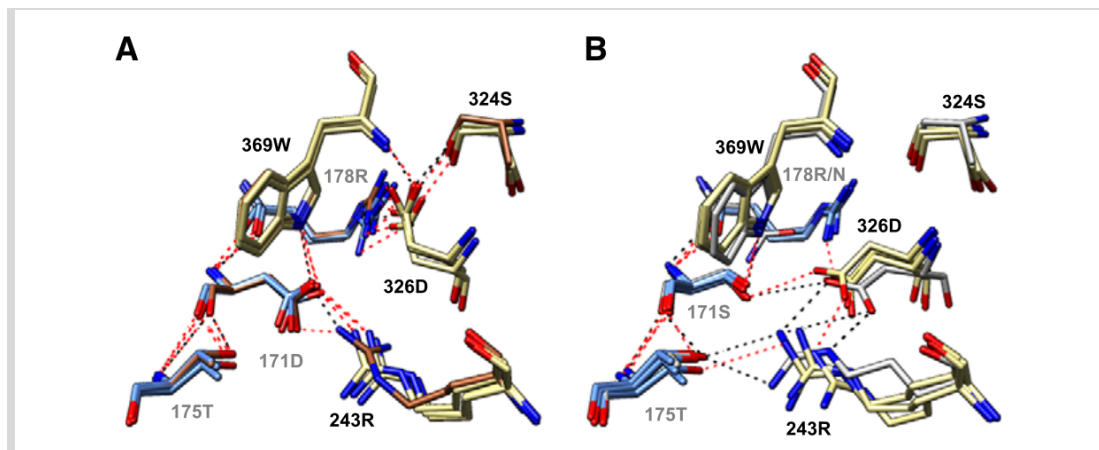


Figure 20: Illustration structurale des possibles conséquences de la flexibilité des chaînes principales en design. Comparaisons des résidus à l'interface avec l'acide aspartique 326 du récepteur. Les résidus du récepteur et du ligand sont représentés, respectivement, en jaune avec annotation noire, et en bleu avec annotation grise. Les lignes en pointillés noir et rouge indiquent, respectivement, les liaisons hydrogène de la structure cristallographique ou des modèles prédits. A) Comparaison entre 3 structures prédites favorisant le résidu 171D et la structure cristallographique, représentée en marron (identifiant PDB: 1A22). B) Comparaison entre 3 structures prédites favorisant la mutation D171S et la structure cristallographique d'un mutant comportant notamment les mutations D171S et R178N, représentée en gris (identifiant PDB: 1KF9). D'après Humphris et Kortemme, 2008.

Des applications de Rosetta pour le re-design ont aussi été documentées pour l'augmentation d'affinité envers des récepteurs. Par exemple, le re-design de l'hormone de croissance humaine (hGH) en vue d'augmenter l'affinité envers son récepteur (hGFR) a été comparé à des résultats de diversité explorée par *phage display* (Humphris et Kortemme, 2008). De façon intéressante, les auteurs ont pu observer que l'enrichissement prédit est globalement corrélé avec les résultats expérimentaux, aboutissant au design d'une banque computationnelle proche de celle observée lors de la maturation d'affinité réalisée en *phage display*. D'après les auteurs, cette étude illustre également l'impact que peut avoir l'introduction de mouvements du squelette peptidique avec, par exemple, une conformation différente du résidu D326 (Figure 20) favorisant la mutation

D171S. Une autre interface impliquant un récepteur a été soumise à un re-design prédictif, en modifiant la surface d'un anticorps monoclonal (Herceptine) fixé au récepteur d'hormones de croissance épidermique humain HER2 (Babor et al., 2011). A partir de la structure cristallographique du complexe, les auteurs ont mis en avant l'impact positif de la modélisation d'une légère flexibilité pour identifier la mutation D98W (générant un gain d'affinité $\times 3$) observée en *phage display*. L'introduction de mouvements plus amples, par relaxation ou simulation de dynamique moléculaire, n'a cependant pas permis d'améliorer les prédictions.

1.4.3.5. *Design de novo*

Constatant les quelques succès d'ancrage moléculaire et de re-design, la communauté d'utilisateurs et contributeurs de Rosetta a validé et amélioré ces algorithmes pour prédire *de novo* des complexes inter-protéiques affins et spécifiques. Les premiers exemples ont essentiellement comporté des mutagenèses computationnelles sur les deux faces des partenaires liés, comme dans le cas du design *de novo* des structures tridimensionnelles, mais également des surfaces d'interaction de protéines ordonnées. Par exemple, de telles approches ont été employées pour modifier des surfaces d'interaction d'hélices et former des cristaux de géométrie P6 (Lanci et MacDermid, 2012) ou d'autres complexes ordonnés (tétra- ou octaédriques) générés par ancrage et design de blocs protéiques (King et al., 2012), tels que représentés sur la **Figure 21**.

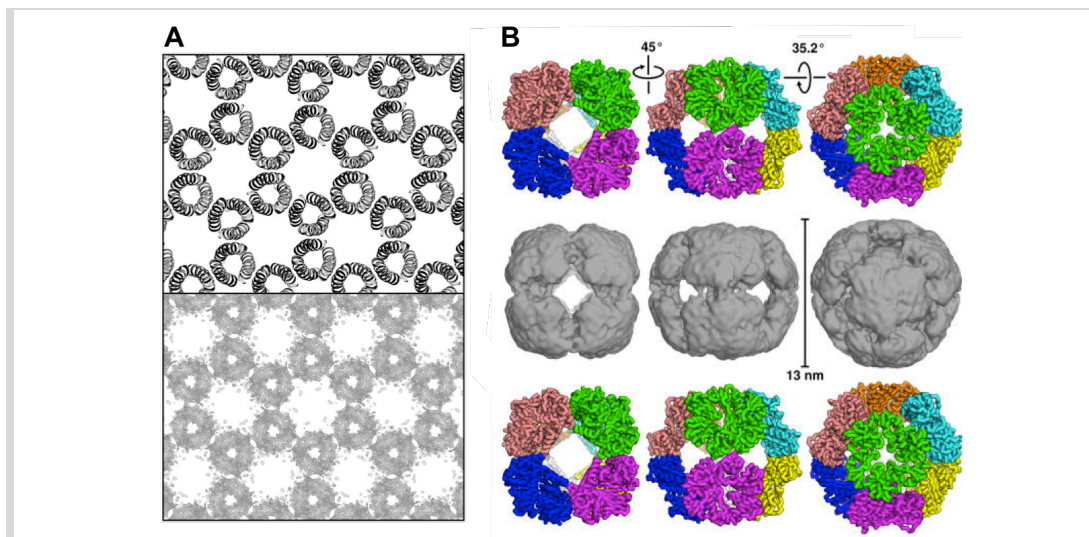
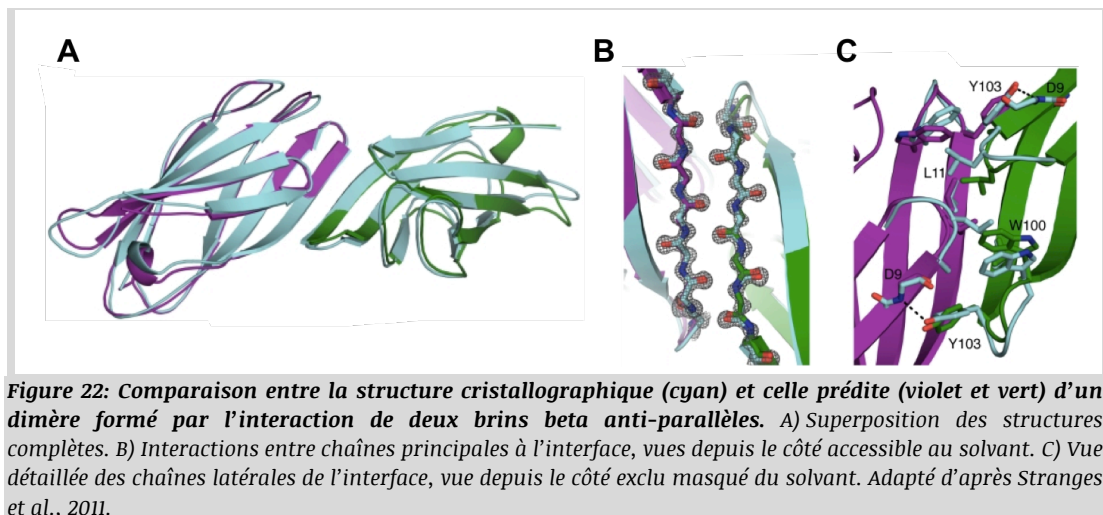


Figure 21: Exemples d'assemblages protéiques réalisés par design de novo. A) Comparaison entre structure prédite (en haut) et densité électronique du cristal d'un complexe de géométrie P6. Adapté d'après Lanci et MacDermid, 2012. B) Comparaison entre structure prédite (en haut), carte de densité par cryomicroscopie électronique (milieu) et structure cristallographique (bas) d'un complexe octaédrique de géométrie R32. Adapté d'après King et al., 2012.

Le design *de novo* a également été appliqué sur des brins beta exposés, mutés pour favoriser la formation d'interactions hydrogènes et ainsi mimer la formation d'un feuillet beta anti-parallèle (Stranges *et al.*, 2011). Grâce à cette stratégie, des monomères ont ainsi été détournés en homodimères symétriques avec une affinité de l'ordre du micromolaire (**Figure 22**). Par ailleurs, l'oligomérisation d'une protéine normalement monomérique a aussi été accomplie à partir du cytochrome cb_{562} , en intégrant des sites de coordination au zinc à l'interface prédite par design computationnel (Salgado *et al.*, 2010). Les auteurs ont ensuite réussi à effectuer le re-design des interfaces dirigées par la coordination du métal pour se dispenser de la présence de zinc dans le maintien de l'interaction (avec des K_D estimés entre 25 et 55 μ M).



D'autres exemples d'applications de Rosetta se sont rapprochés des contraintes liées au design de protéine d'affinité envers une cible non altérée, en effectuant le design *de novo* d'interfaces via l'introduction de mutations à la surface d'un des partenaires uniquement. C'est notamment le cas du design d'un mimétique du peptide non conformé à l'extrémité C-terminale du régulateur RGS14 (également appelé motif GoLoco) se liant à une sous-unité alpha de protéine G ($G\alpha_i$), impliquant la prédiction de la conformation du peptide généré et de son site de fixation (Sammond *et al.*, 2011). Dans cette étude, les auteurs ont ainsi ciblé une gouttière protéique hydrophobe avec un peptide de 16 acides aminés adoptant une structure en hélice lors de sa fixation, mimant ainsi les 12 résidus du motif GoLoco naturellement liés (**Figure 23**). Dans cette approche, les auteurs ont effectivement réussi à générer un peptide spécifique via des

interactions hydrophobes (évitant ainsi la modélisation de l'établissement de liaisons hydrogène correctement orientées entre les chaînes principales ou latérales des résidus au sein de l'interface) avec un squelette peptidique adoptant une conformation proche de celle prédite (RMSD de la chaîne principale de 1,1 Å).

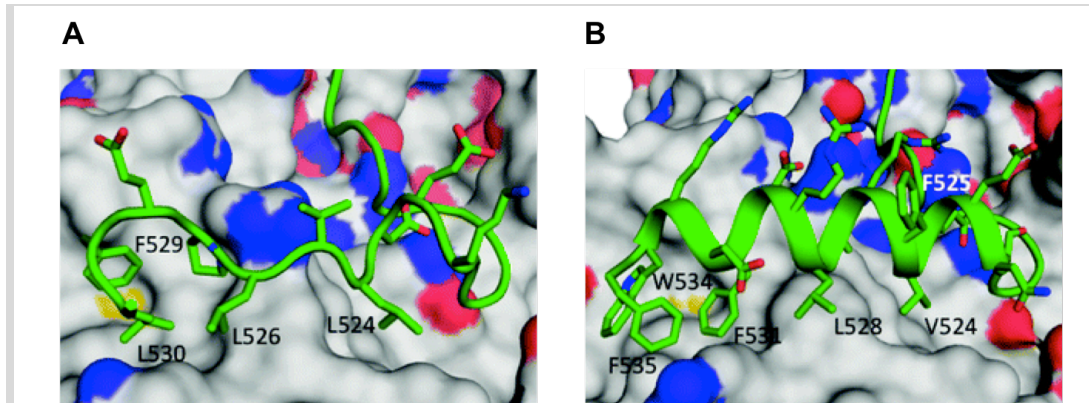


Figure 23: Design d'un mimétique du motif GoLoco. Les peptides sauvage et généré *de novo* sont représentés en cartoon vert. La surface de $G\alpha_{11}$ est figurée en gris. Les azotes et oxygènes des chaînes latérales sont colorés, respectivement, en bleu et rouge. A) Structure cristallographique du peptide motif GoLoco lié à $G\alpha_{11}$. B) Modèle du mimétique après design, tel que sa liaison à $G\alpha_{11}$ a été prédite. Adapté d'après Sammond et al., 2011.

Un exemple remarquable de design *de novo* opéré par Rosetta est celui de la formation d'une interaction affine entre les protéines Pdar et Prb, ayant mené à la découverte de variants avec un K_D de 130 nM après design, puis 180 pM après maturation d'affinité *in vitro* (Karanicolas et al., 2011). Ces affinités, situées dans une gamme compatible avec des applications biotechnologiques et thérapeutiques, ont été obtenues après recherche de complémentarité spatiale (par ancrage moléculaire global puis local) puis design de l'interface sur chaque partenaire pour établir une liaison hydrogène entourée d'interactions hydrophobes compactes (**Figure 24A et B**). Ce résultat est cependant mitigé par une rotation induite lors de la maturation d'affinité. En effet, le résidu Prb(R89) serait décalé vers le résidu Pdar(D83) après mutation d'une asparagine, tandis que le résidu Prb(N83) serait rapproché du centre de l'interface après mutation d'un acide aspartique (**Figure 24C et D**). Malgré cette modification d'orientation du complexe, l'interface ciblée lors du design *de novo* reste celle désirée, ce qui est tout à fait satisfaisant pour de nombreuses applications (dont le développement de protéines neutralisantes).

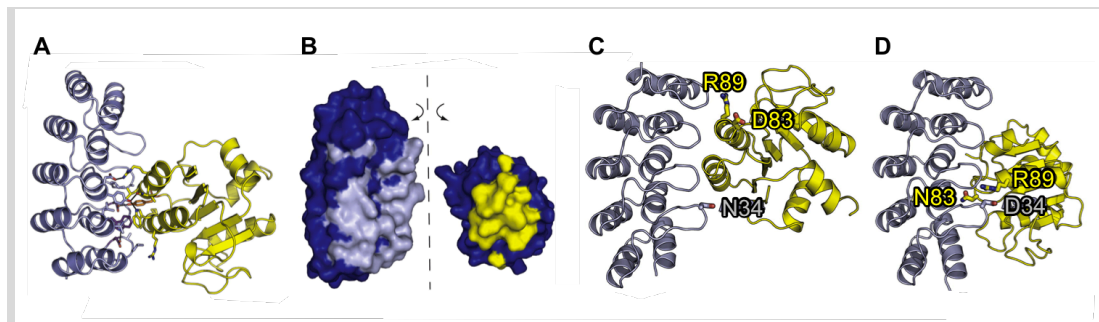


Figure 24: Design de novo du complexe Pdar-Prb et rotation induite lors de la maturation d'affinité in vitro. Les structures en cartoon des variants de Pdar et Prb sont colorées, respectivement, en gris et en jaune. A) Modèle du complexe Pdar-Prb après design, avec les motifs centraux figurés en bâtonnets. B) Modèle du complexe après design, séparé pour révéler les surfaces d'interactions de Pdar (gris) et Prb (jaune) ayant été randomisées lors du design. C) Modèle du complexe Pdar-Prb après design, avec les résidus Pdar(N34), Prb(R89) et Prb(D83) figurés en bâtonnets. D) Modèle du complexe entre Pdar et Prb pivotée de 180°, en raison des mutations Pdar(N34D) et Prb(D83N) sélectionnées lors de la maturation d'affinité. Adapté d'après Karanicolas et al., 2011.

Un autre exemple, plus proche de nos attentes méthodologiques, a consisté à effectuer l'ancrage moléculaire d'une charpente protéique en hélices sur des régions définies à la surface du domaine kinase PAK1, puis à effectuer le design de l'interface en introduisant des mutations uniquement à la surface de la charpente (Jha et al., 2010). Pour cela, des simulations d'ancrage moléculaire (sans flexibilité des chaînes principales) ont été réalisées, suivies de design et de minimisation des chaînes latérales et principales pour obtenir les interfaces de plus faible énergie. Les auteurs ont également décrit l'apport de simulations en dynamique moléculaire pour déterminer le choix de 6 charpentes mutées à caractériser expérimentalement, aboutissant à 2 protéines agrégées et 3 charpentes capables de fixer PAK1 avec des affinités de 100 à 330 μM .

En 2010, une étude menée par le groupe de Kuhlman a comparé les approches de design computationnel (capable d'explorer un large espace de séquence) et de sélection (non limitée par l'exploration de l'espace conformationnel des variants) avec criblage par complémentation (Guntas et al., 2010). Ainsi, ils ont effectué le design de mutants d'une ubiquitine ligase vers un partenaire non naturel en comparant une banque naïve de 13 positions (20 acides aminés autorisés par position), une banque semi-dirigée prédite pour éliminer les variants les plus instables (indépendamment de leur capacité prédite à fixer leur cible), et une banque conçue en fonction des prédictions d'ancrage moléculaire et de design. Après 4 tours de sélection, les auteurs rapportent avoir identifié plusieurs protéines affines ($K_D < 100 \text{ nM}$) avec la banque issue de design, mais également avec celle semi-dirigée. Ces banques ont par ailleurs montré un

enrichissement 30 fois plus important que la banque naïve lors du premier tour de sélection (qui a abouti à des affinités avec des $K_D > 50 \mu\text{M}$ après les 4 tours de sélection). Il est intéressant de noter que l'optimisation des banques de protéines d'affinité présente un effet positif sur les performances de sélection, y compris quand cette optimisation se fait uniquement sur la stabilité des variants.

Pour finir, nous détaillerons les résultats d'une étude ayant généré avec succès des protéines affines d'un épitope conservé de l'hémagglutinine virale (dont l'importance stratégique a été soulignée dans la section "I.4.1. Protéines de liaison et reconnaissance moléculaire spécifique"). Après design *de novo*, deux protéines ont été identifiées avec des K_D supérieurs au micromolaire (Fleishman *et al.*, 2011). Leur maturation d'affinité par *yeast display* a ensuite permis de réduire leur K_D à la dizaine de nanomolaire. La première de ces protéines, HB80, a démontré un effet inhibiteur sur un changement conformationnel pH-dépendant de l'hémagglutinine. La seconde, HB36, a été co-cristallisée avec sa cible et a été remarquablement prédite en matière d'orientation lors de cette étude (**Figure 25A et B**). La démarche adoptée par les auteurs pour fournir ces résultats est résumée sur la **Figure 25C**. En bref, un épitope défini a été ciblé sur la surface de l'hémagglutinine (épitope CR6261) puis des charpentes protéiques variées ont été criblées pour leur complémentarité géométrique avec cet épitope (en tenant compte de contraintes de compatibilité de points chauds sur les deux partenaires). Ensuite, des simulations d'ancrage moléculaire ont été réalisées avant d'intégrer des points chauds sur les charpentes, puis de réaliser le design des autres résidus des charpentes situés à l'interface. Enfin, les meilleurs résultats de design ont été validés expérimentalement après maturation d'affinité.

La série d'applications de Rosetta que nous avons décrite ici dresse le tableau des résultats ponctuellement générés par ses utilisateurs. D'une part, ceci étaye la capacité d'amélioration de cette suite logicielle ainsi que la possibilité d'appliquer ses méthodes à différents niveaux de l'ingénierie des protéines. D'autre part, ces résultats prouvent que Rosetta constitue une méthode suffisamment efficace pour justifier son utilisation dans notre approche de design *de novo* de Nanofitines. Dans le chapitre "Chapitre III: Design *de novo* de Nanofitine spécifique", nous

décrivons une stratégie d'implémentation de Rosetta comme complément de la sélection *in vitro* de Nanofitines, notamment pour l'identification d'épitopes précis puis pour le design ultérieur de Nanofitines spécifiques à ces interfaces identifiées, dans une preuve de concept articulée autour de la GFP.

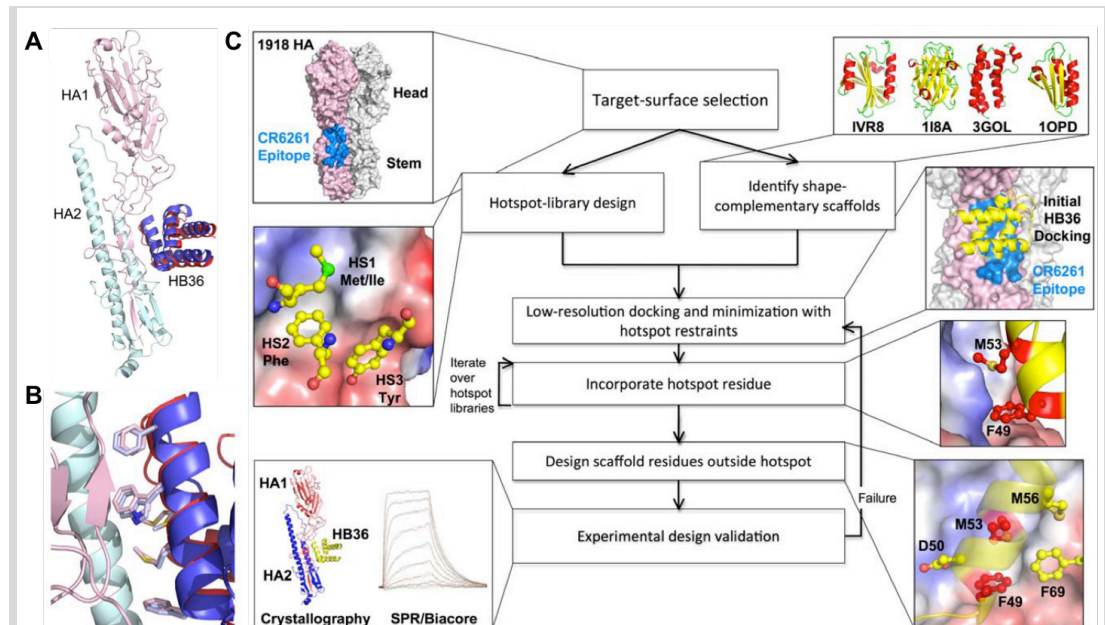
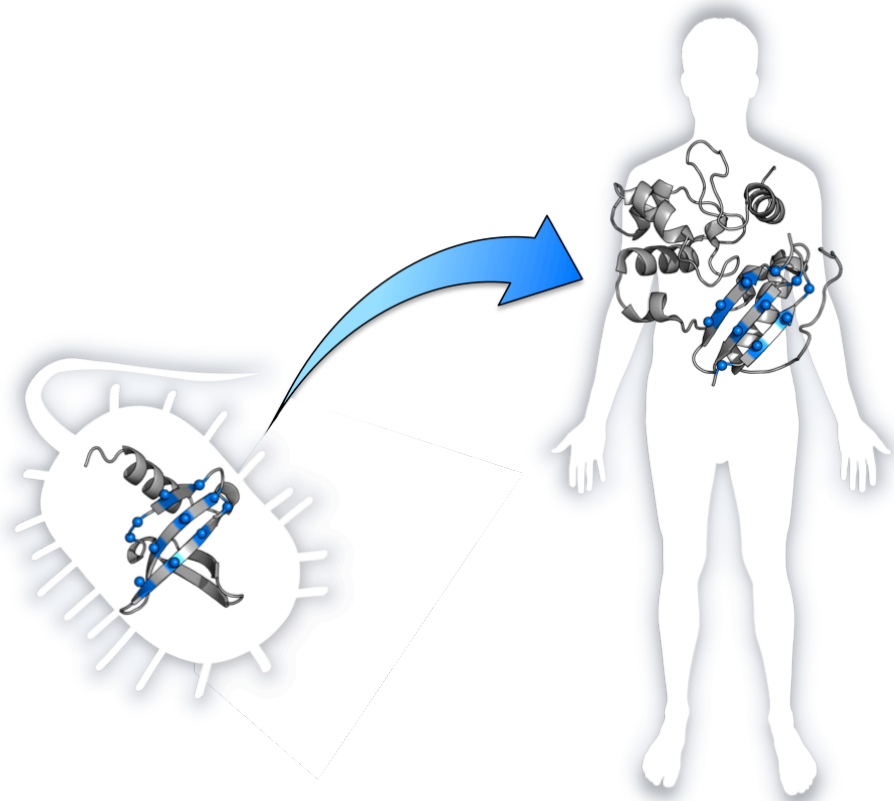


Figure 25: Design de protéine spécifique et affine d'une région conservée de l'hémagglutinine. A) Structure cristallographique du complexe entre l'hémagglutinine H1 (sous-unités HA1 et HA2 représentées en rose et cyan, respectivement) et un variant mûré de HB36 (rouge), superposée au modèle prédit. La structure prédite de HB36 est représentée en bleu, tandis que celle de l'hémagglutinine n'est pas figurée. B) Gros plan sur l'interface représentée en A), avec ajout des représentations en bâtonnets de résidus de HB36 (bleu clair) et de son variant mûré co-cristallisé (rose). C) Logigramme des étapes clés de la démarche adoptée pour la découverte des anti-hémagglutinine HB36 et HB80. Adapté d'après Fleishman et al., 2011.

Chapitre II: Humanisation de Nanofitines par greffe de domaine



Chapitre II: Humanisation de Nanofitines par greffe de domaine

II.1. Stratégie d'humanisation des Nanofitines

II.1.1. Des domaines de liaison exposés sur le feuillet beta de Sac7d

Artificiellement dérivées de la protéine Sac7d, les Nanofitines trouvent leur origine dans une protéine archéobactérienne, soulevant un questionnement sur le risque d'immunogénicité de ces protéines de liaison en vue de leur utilisation dans des applications thérapeutiques. Bien qu'il soit possible d'argumenter que l'origine humaine d'une protéine n'est pas le gage d'une absence d'immunogénicité et que les Nanofitines présentent des propriétés physicochimiques qui tendent à minimiser le déclenchement de réactions immunitaires (grâce à la combinaison d'une petite taille, d'une extrême stabilité et du maintien d'un profil monomérique à des concentrations > 100 mg/ml), Affilogic a souhaité investir dans un programme de recherche visant à explorer une opportunité d'humanisation des Nanofitines. Cette démarche s'inscrit dans une volonté d'étendre le spectre des possibilités offertes par la découverte de molécules d'affinité dérivées de Sac7d par *ribosome display*, en offrant l'éventualité d'exploiter la charpente robuste des Nanofitines ou de tirer parti de leur site de liaison en le greffant sur une protéine hôte bénéficiant d'autres propriétés intrinsèques. Dans ce cadre, la démarche d'humanisation que nous avons adoptée réside donc dans la greffe des 11 résidus sélectionnés par *ribosome display* à la surface des Nanofitines vers des protéines humaines arborant un environnement jugé compatible. La preuve de concept de l'humanisation de Nanofitines discutée dans ce chapitre a été réalisée avec des interactions modèles entre la protéine de fluorescence verte (GFP) et le feuillet beta rigide des Nanofitines, à partir de deux sites de fixation spécifiques de la GFP transférés sur trois charpentes humaines ainsi que sur Sso7d, un homologue structural des Nanofitines.

A l'heure actuelle, plus de 40 projets ont été menés avec succès à Affilogic en mettant à profit des Nanofitines dont la surface d'interaction mutée est majoritairement exposée à la surface du second feuillet beta de Sac7d. Partant de cette observation, nous avons donc cherché à mettre en place le transfert de résidus fonctionnels de Nanofitines vers des protéines humaines ayant un

feuillet beta compatible avec celui sélectionné par *ribosome display* sur Sac7d. Les exemples que nous avons évoqués précédemment (dans la section “I.3. Transfert de fonction et stratégie d’humanisation”) ont essentiellement été validés lors du transfert de boucles labiles, ou d’hélices alpha. L’approche de transfert de brins beta consécutifs se démarque de ces illustrations et demande donc d’identifier les bénéfices et contraintes de cette stratégie. Le principal avantage que nous avons anticipé réside dans la réduction de la labilité en comparaison avec le transfert de boucles, généralisé lors de l’humanisation des anticorps. En effet, nous avons déjà abordé la difficulté de greffer certaines boucles dans la section “I.3.3.2. Transferts de boucles: miniaturisation de charpentes”. De plus, il est généralement admis que les boucles présentent des challenges de modélisation accrus par leur longueur, ce qui est principalement le cas de la boucle CDR-H3 (*Kuroda et al., 2012*), bien que l’insertion de longues boucles semble pourtant avantageuse sur le plan énergétique (*Scalley-Kim et al., 2003*). Il semblerait que ces défis méthodologiques soient à l’avantage de la prédiction de sites accepteurs du feuillet beta des Nanofitines, plus rigide et donc plus facile à modéliser.

II.1.2. Enseignements du transfert d'un site de Nanofitine entre feuillet et boucles

Le postulat précédent est conforté par des expériences préliminaires de transfert d’un site de fixation aux immunoglobulines G découvert par mutation aléatoire de deux boucles et d’une portion du second feuillet beta de Sac7d (*Béhar et al., 2014*). Après greffe des résidus mutés sur l’homologue Sso7d (dont les mutations sont essentiellement situées aux extrémités N- et C-terminales de la protéine, soit la face opposée au feuillet), la Nanofitine a bénéficié d’une amélioration de profil de stabilité en condition alcaline, mais sa température de fusion a chuté (T_m passant de 80,4°C à 74,2°C) alors que Sso7d est plus thermostable que Sac7d (T_m de 100,4 et 90,2°C respectivement). Plus important encore, la Nanofitine a vu son affinité de 34 nM réduite à l’ordre du micromolaire lors du transfert, alors que nous pouvions anticiper une plus forte tolérance des boucles vis-à-vis des modifications induites dans leur nouvel environnement (compte tenu de leur flexibilité). Il a d’ailleurs été suggéré que les protéines thermostables sont plus rigides que les protéines d’origine mésophile (*Vihinen, 1987; Reetz et al., 2006*), et que les

protéines les plus stables sont plus tolérantes aux mutations et promeuvent l'évolvabilité de leurs séquences et fonctions (Bloom *et al.*, 2006). De ce fait, le transfert d'un site de fixation rigide (généré sur le feuillet beta d'une Nanofitine) semble une alternative prometteuse. Cette rigidité devrait également permettre de modéliser plus aisément la conformation adoptée par la chaîne principale des résidus fonctionnels. En absence de motifs étendus sur des structures secondaires plus labiles (comme la première boucle de Sac7d décrite lors des premières conceptions de banques de Nanofitines), cette démarche pourrait finalement faciliter l'identification d'environnements compatibles au transfert.

II.1.3. Transfert du feuillet anti-GFP de Nanofitines

Compte tenu des sites d'interactions privilégiés lors de la sélection de Nanofitines ainsi que de l'origine archéobactérienne de la charpente de Sac7d, nous avons proposé une stratégie d'humanisation des Nanofitines par greffe de leur feuillet beta sur des protéines d'origine humaine. Lors de l'évaluation de notre démarche (résumée **Figure 26**), nous avons identifié la conformation adoptée par le squelette peptidique des brins formant le feuillet beta de Sac7d, ce qui autorise potentiellement la découverte de feuillets discontinus ou inversés. Le motif a ensuite été recherché de manière itérative dans les structures résolues de protéines humaines de la PDB pour découvrir les receveurs de l'interface issue de Nanofitine. Ces protéines, après greffe des résidus sélectionnés par *ribosome display*, ont été exprimées puis caractérisées pour évaluer leur nouvelle capacité de fixation.

A priori plus aisé à modéliser, le choix de ce transfert d'interface présente tout de même des particularités par rapport aux autres démarches portées à notre connaissance. En effet, nous cherchons à greffer des résidus d'une protéine simple et hyperstable vers des protéines non homologues probablement plus complexes et moins robustes (sur la base de leur environnement naturel) afin de pouvoir humaniser les Nanofitines. Par ailleurs, nous ne connaissons pas la capacité des protéines humaines sélectionnées à être exprimées *in vitro* ou à supporter l'introduction de mutations additionnelles à celles du transfert. Par manque de maîtrise des charpentes acceptrices, le défi de notre stratégie a donc été rehaussé car nous n'avions pas la

possibilité matérielle de réaliser une étape supplémentaire de maturation d'affinité *in vitro*. Nous avons donc assumé le risque de ne pas pouvoir réellement constater les résultats de cette stratégie en obtenant des variants greffés en une étape unique sans voie d'amélioration.

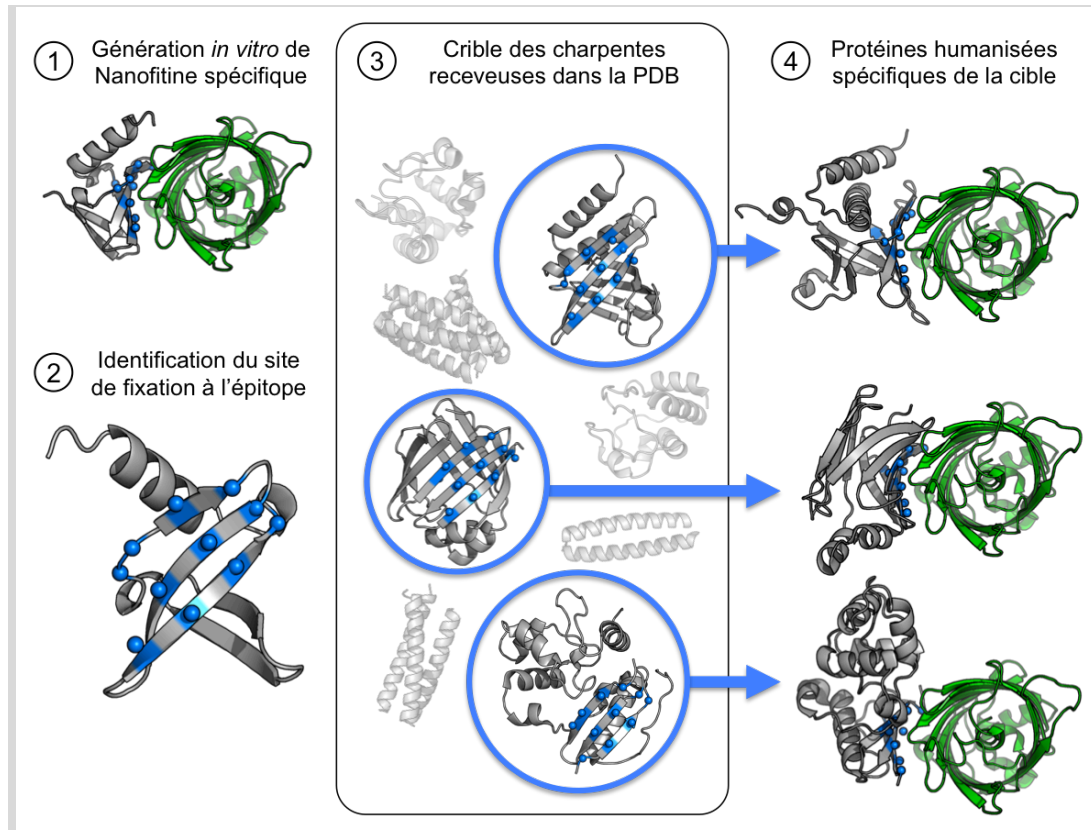


Figure 26: Représentation schématique de la stratégie d'humanisation des Nanofitines. Des Nanofitines sont sélectionnées par ribosome display contre une cible d'intérêt (comme la GFP) par randomisation des résidus exposés sur le second feuillet beta de Sac7d (étape 1). Les coordonnées spatiales des résidus randomisés sont identifiées (étape 2) pour être recherchées dans les structures de protéines humaines disponibles dans la PDB (étape 3). Finalement, les protéines présentant un environnement proche de celui sur le feuillet des Nanofitines (conformation du squelette peptidique, orientation des chaînes latérales et accessibilité) sont greffées avec les résidus sélectionnés *in vitro* pour obtenir l'activité de fixation spécifique de la cible (étape 4). Les charpentes de Sac7d et des potentielles protéines humanisées sont représentées en gris, avec les carbones alpha des résidus recherchés figurés en sphères bleues. La cible est colorée en vert.

Malgré des topologies et surfaces des sites à transférer différentes, nous avons remarqué que des méthodes similaires de recherche des protéines receveuses au sein de la PDB ont été décrites dans la littérature. C'est notamment le cas de STAMPS/RASMOT-3D PRO, dont un exemple récent a permis de rechercher un motif des 4 résidus fonctionnels de l'inhibiteur tissulaire de métalloprotéases TIMP-2 dans de petites protéines de la PDB (Tlatli *et al.*, 2013). Grâce à une recherche d'orientation spatiale des carbones alpha et beta, autorisant l'identification de motifs discontinus, dix charpentes receveuses ont été identifiées et greffées des quatre acides aminés. Les auteurs ont rapporté des affinités avant optimisation allant de 0,45 μM à 450 μM pour

différentes métalloprotéases, mais également des différences de géométrie et de stabilité des variants générés, que nous espérons limiter dans le cas du transfert d'un feuillet beta rigide. L'algorithme utilisé dans cette étude avait été décrit précédemment et appliqué pour générer des mini-protéines dirigées contre un canal potassique, aboutissant à 3 candidats avec des K_D entre 0,5 et 1,6 μM après transfert des 4 résidus fonctionnels (Magis et al., 2006). Une recherche similaire de motifs dans la PDB, également basée sur l'orientation spatiale des carbones alpha et beta, a aussi été décrite dans une étude visant à greffer les 3 résidus clés de l'érythropoïétine (EPO) humaine vers un domaine de pleckstrine de rat (S. Liu et al., 2007). Ces travaux ont montré que les variants générés ont gagné une activité de fixation du récepteur à l'érythropoïétine (K_D entre 24 et 500 nM), après introduction d'une à quatre mutations.

Les résultats de notre approche, réalisée à partir de deux sites à greffer sur trois protéines humaines et l'homologue Sso7d, sont discutés dans la suite de ce chapitre. Nous y décrivons la mise en œuvre de notre stratégie d'humanisation à partir de Nanofitines anti-GFP découvertes en amont des travaux présentés dans cette thèse, mutées uniquement sur leur second feuillet. Via cette démonstration, notre objectif est d'étendre la surface transférée avec succès entre protéines non homologues afin d'obtenir des Nanofitines humanisées. En cas de succès, cette démarche placerait les Nanofitines actuelles (facilement sélectionnées *in vitro* à partir de Sac7d) comme support de découverte de sites de liaison et comme précurseurs de protéines d'affinité humanisées à vocation thérapeutique.

II.2. Découverte de Nanofitines anti-GFP interagissant via leur feuillet

II.2.1. Choix de la banque de variants de Nanofitines

La famille protéique de Sac7d, comprenant des protéines homologues telles que Sso7d, a démontré une flexibilité élevée et une importante tolérance en tant que charpente pour la conception de banques de mutants. Ces propriétés permettent la génération de molécules de fixation avec une forte affinité et spécificité envers diverses cibles, via la mutagenèse aléatoire des Nanofitines sur leur feuillet beta (Mouratou et al., 2007) mais également leurs boucles, avec (Correa et al., 2014) ou sans extension (Béhar et al., 2013). Ces précédentes études mettent en

exergue la stabilité de cette nouvelle charpente protéique à l'égard de l'introduction d'un nombre conséquent de mutations, pouvant couvrir au moins 27% des positions disponibles sur Sac7d (Béhar *et al.*, 2013). De telles mutations ont été introduites avec succès à l'aide de différentes banques de variants, mais également grâce au transfert de domaine entre membres de la famille des OB-fold (Béhar *et al.*, 2014).

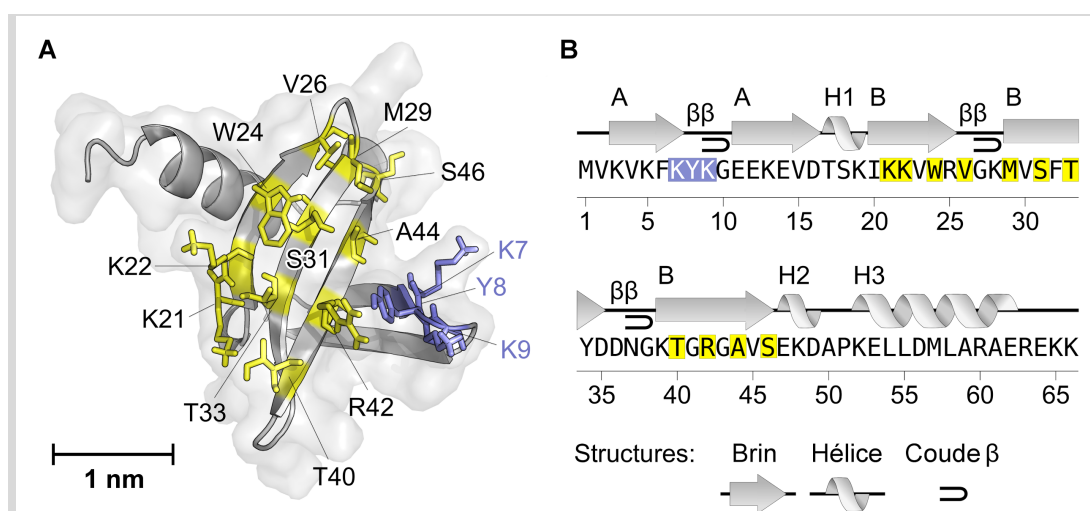


Figure 27: Conception de banques de Nanofitines à partir de Sac7d. A) Représentation schématique de Sac7d sauvage (Identifiant PDB: 1AZP). B) Diagramme de la structure secondaire de Sac7d sauvage. Les hélices sont annotées H1, H2 et H3 et les brins sont par leur feuillet A et B, de l'extrémité N-terminale à l'extrémité C-terminale. Les motifs en coude beta sont indiqués par β . Les résidus rendus aléatoires dans la banque utilisée lors de la génération de Nanofitines anti-GFP sont représentés par des bâtonnets jaunes (A) ou encadrés jaunes (B). Les résidus de la première boucle rendus aléatoires dans la première conception de banque sont représentés par des bâtonnets violets (A) ou encadrés violets (B).

Lors de la première conception de banque de Nanofitines, les acides aminés impliqués dans l'interaction avec l'ADN double brin, ligand naturel de Sac7d, ont été rendus aléatoires pour rediriger cette capacité de fixation vers de nouvelles cibles protéiques. Cette première banque a consisté à muter aléatoirement 14 résidus situés dans la première boucle et le second feuillet beta de la charpente (Mouratou *et al.*, 2007; **Figure 27**). Une telle banque couvre donc une diversité théorique de plus de $1,6 \times 10^{18}$ variants, ce qui dépasse la capacité maximale de criblage estimée pour les différentes technologies de sélection *in vitro* (comprise entre 10^{10} pour les techniques impliquant la transformation de cellules vivantes comme le *phage display* (Smith, 1985; Winter *et al.*, 1994) ou le *yeast display* (Chien *et al.*, 1991; Fields et Song, 1989), et 10^{14} pour les techniques purement *in vitro* comme le *ribosome display* (Schaffitzel *et al.*, 2001)). Il a logiquement semblé nécessaire d'utiliser des banques dont la diversité pouvait être intégralement explorée par le

procédé de *ribosome display*, notamment pour augmenter la reproductibilité de la sélection de variants de Sac7d. Des banques alternatives ont ainsi été générées au sein d’Affilogic, réduisant le nombre de mutations introduites à 11 positions ou moins, soit une diversité maximale de 2×10^{14} variants. Par ailleurs, une diversité réduite a déjà été prouvée suffisante pour isoler des protéines d’affinité (allant de 34 nM au μM) envers diverses cibles à partir de la mutagenèse de 10 positions de structures *OB-fold*, comme réalisé précédemment par Pecorari et ses collaborateurs par *ribosome display* sur la charpente Sac7d (Béhar et al., 2013), ou par Rao et ses collaborateurs par *yeast display* sur la charpente Sso7d (Gera et al., 2012, 2011; Hussain et al., 2013).

Nous avons focalisé l’approche d’humanisation des Nanofitines sur des clones préalablement obtenus à partir d’une de ces banques réduites, se résumant à l’introduction de 11 mutations aléatoires dans le second feuillet beta de Sac7d (**Figure 27**). Au-delà des considérations de reproductibilité du procédé de sélection, le choix de cette banque recouvrant le feuillet beta des Nanofitines a été réalisé dans le but de générer des sites de fixation dont la structure puisse présenter une labilité réduite, notamment par rapport aux structures secondaires en boucles ou en coudes. Cet aspect apparaît crucial dans la conception de notre démarche d’humanisation des Nanofitines, puisque la rigidité de l’agencement spatial des acides aminés fonctionnels à transférer sur d’autres charpentes rend ces résidus plus facilement identifiables dans les représentations relativement figées que sont les structures cristallographiques. Pour réaliser la démonstration de notre approche, nous avons donc isolé des Nanofitines spécifiques de la GFP lors d’une sélection par *ribosome display* (**Figure 28**) en aval des travaux de transfert, en utilisant la banque randomisant 11 positions à la surface du feuillet beta de Sac7d. Par ailleurs, nous avons identifié des charpentes protéiques humaines dont les structures cristallographiques suggèrent la présence d’un feuillet compatible au transfert des résidus sélectionnés, car similaire à celui muté lors de la sélection *in vitro* (tel que nous le décrirons ultérieurement dans la section “II.3. Charpentes protéiques humaines hôtes de feuillet anti-GFP”).

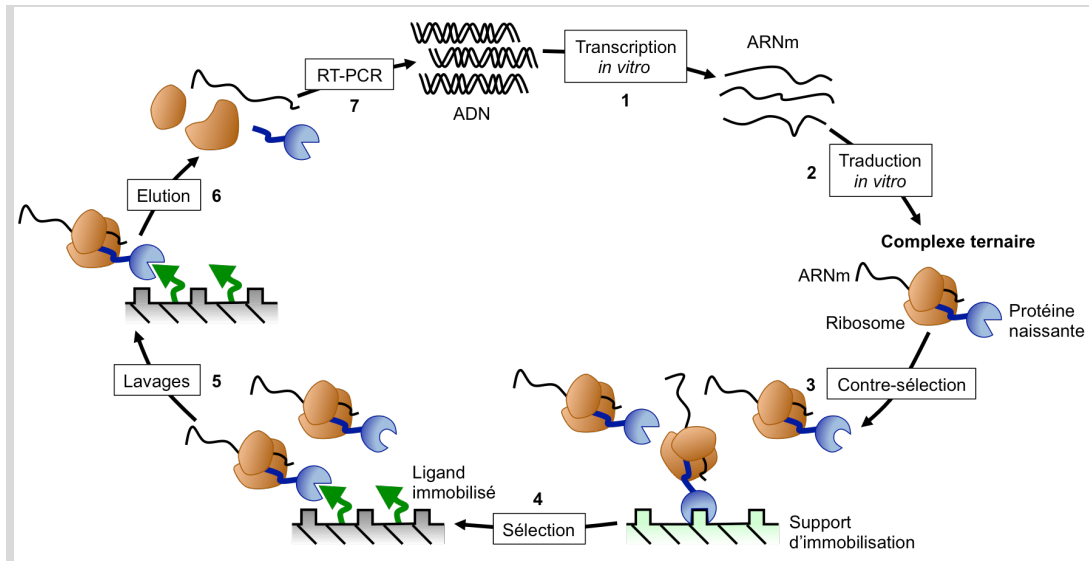


Figure 28: Principe du ribosome display. Représentation schématique d'un tour de sélection par ribosome display. 1) Une banque d'ADN avec un cadre de lecture sans codon stop est transcrite in vitro. 2) L'ARN messager est purifié puis traduit in vitro, formant un complexe ribosomal composé de l'ARNm, du ribosome et de la protéine néo-synthétisée séparée du ribosome par un espaceur. 3) Les complexes ternaires exposant des Nanofitines affines pour l'environnement de sélection sont éliminées lors d'une étape de pré-panning. 4) Les complexes ternaires exposant des Nanofitines affines pour le ligand immobilisé sont capturées par le ligand immobilisé sur support. 5) Les complexes ternaires exposant des Nanofitines non liées au ligand immobilisé sont éliminées lors de lavages. 6) Les complexes ribosomaux liés sont élués et les ARNm purifiés. 7) Les ARNm purifiés sont rétro-transcrits et amplifiés par PCR, pour être engagés dans un nouveau tour de sélection.

II.2.2. Sélection des Nanofitines anti-GFP par ribosome display

Après 4 tours de sélection par ribosome display, un criblage a été réalisé par ELISA sur 96 clones isolés dont 80 présentaient une fixation importante et spécifique à la GFP (avec une augmentation de signal d'au moins 10 fois lors de l'ajout de GFP, **Figure 29**).

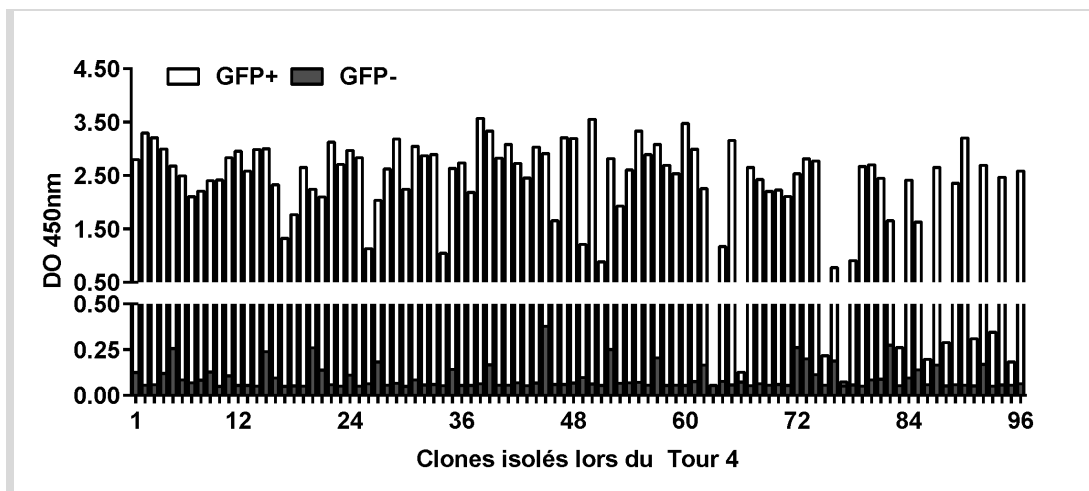


Figure 29: Criblage des Nanofitines anti-GFP par ELISA au quatrième tour de sélection. Criblage réalisé en présence (barres blanches) ou absence (barres grises superposées) de StrepTagII-GFP immobilisée.

Afin de réduire l'effort de criblage et restreindre la diversité observée aux Nanofitines les plus affines, la sélection a été intensifiée pendant 2 tours supplémentaires avec une concentration

décroissante de protéine cible. Après le sixième tour de sélection, un second criblage ELISA, réalisé sur 105 extraits cellulaires dilués, a permis d'identifier plus de 50 clones isolés présentant un signal spécifique en présence de GFP avec une intensité au moins 10 fois plus grande que celle du bruit de fond observée en absence de GFP (**Figure 30**).

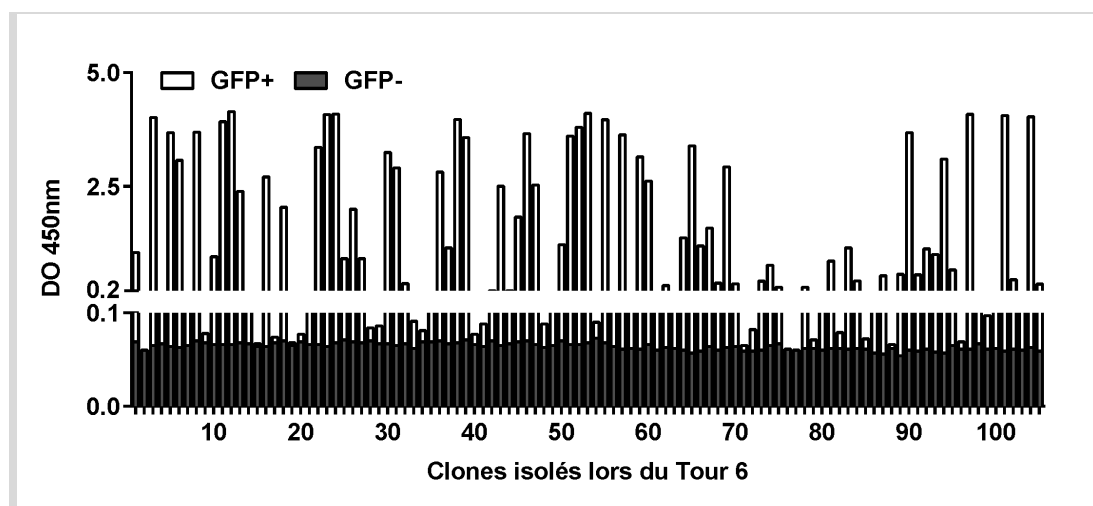


Figure 30: Criblage des Nanofitines anti-GFP par ELISA au sixième tour de sélection. Criblage réalisé en présence (barres blanches) ou absence (barres grises superposées) de StrepTagII-GFP immobilisée.

II.2.3. Choix de 2 domaines de Nanofitines anti-GFP à transférer

II.2.3.1. Critères de choix de Nanofitines anti-GFP

Le procédé de sélection par *ribosome display*, aujourd'hui réalisé en routine pour la génération de variants de Sac7d, semble donc avoir permis de produire une population de Nanofitines spécifiques de la GFP grâce à l'introduction de mutations dans leur feuillet beta. Cette spécificité d'interaction a été mise en évidence lors du criblage réalisé par ELISA, révélant plusieurs dizaines de clones comme potentiels candidats au transfert de charpente.

Afin de maximiser les chances de réussite de cette approche d'humanisation, nous avons trié ces Nanofitines avec comme objectif d'identifier les protéines spécifiques de la GFP présentant les meilleures affinités. Idéalement, nous visons des constantes de dissociation à l'équilibre (K_D) inférieures à 10 nM (sur la base des affinités déjà déterminées de Nanofitines précédemment découvertes), ce qui autoriserait une détection de la capacité de fixation à la GFP malgré d'éventuelles pertes d'affinité.

Enfin, nous avons cherché à ne pas limiter cette stratégie d'humanisation au ciblage d'un épitope unique, diversifiant ainsi les environnements propres aux interfaces protéine-protéine à transférer (notamment en matière d'accessibilité, de charges, etc.). Un criblage a donc été réalisé afin d'isoler des Nanofitines pouvant cibler des régions distinctes à la surface de la GFP, tout en satisfaisant les critères de spécificité et d'affinité.

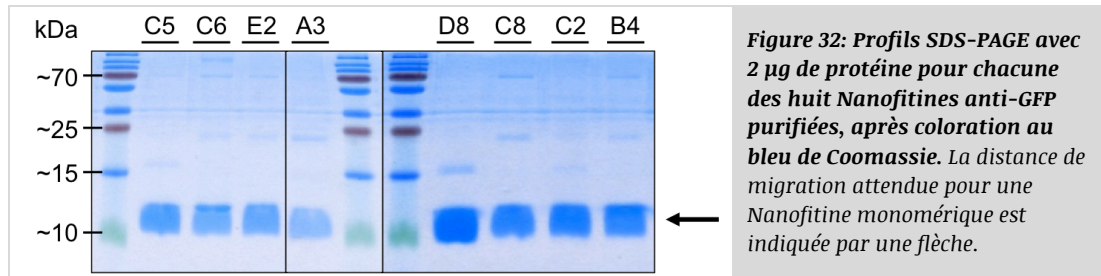
II.2.3.2. Expression et purification des Nanofitines

Sur la base du séquençage des 22 meilleurs clones positifs en ELISA, 8 Nanofitines ont été choisies pour être produites par fermentation bactérienne en erlenmeyer, puis purifiées (**Figure 31**). En plus de confirmer leurs propriétés déduites à partir du criblage des lysats bactériens, la caractérisation des Nanofitines purifiées avait également pour but d'identifier des sites de fixations ciblant des régions distinctes de la GFP.

	1	10	20	30	40	50	60
D8	MVKVKFKYKGEEKEVDTSKI	EMVTR	I GK S VYFH	YDDNGK	W CH Q V L	EKDAPKELL	DMLARAEREKK
C5	MVKVKFKYKGEEKEVDTSKI	N F V TR V G K L V L	F H	YDDNGK	W G V G H V P	EKDAPKELL	DMLARAEREKK
C2	MVKVKFKYKGEEKEVDTSKI	W L V TR F G K S V A	F H	YDDNGK	W G T G V S	EKDAPKELL	DMLARAEREKK
B4	MVKVKFKYKGEEKEVDTSKI	Q L V TR F G K L V T	F H	YDDSGK	W G T G A V Q	EKDAPKELL	DMLARAEREKK
C6	MVKVKFKYKGEEKEVDTSKI	T L V VR F G K L V T	F H	YDDNGK	W G T G A V Q	EKDAPKELL	DMLARAEREKK
A3	MVKVKFKYKGEEKEVDTSKI	E L V VR F G K L V T	F H	YDDNGK	I G V G L V L	EKDAPKELL	DMLARAEREKK
C8	MVKVK Y KYKGEEKEVDTSKI	W A V G R L G K E V	T F	G Y DDNGK	L G A G R V T	EKDAPKELL	DMLARAEREKK
E2	MVKVKFKYKGEEKEVDTSKI	T S V H R L	G K Y V P F V	YDDNGK	T G T G W V D	EKDAPKELL	DMLARAEREKK

Figure 31: Séquences des Nanofitines anti-GFP purifiées et caractérisées. Les mutations par rapport à la séquence de Sac7d sauvage sont soulignées.

Après un unique passage de chromatographie d'affinité sur 300 µL de résine de nickel immobilisé (IMAC), nous avons obtenu de 350 µg à 8,58 mg de chaque Nanofitine purifiée, comme déterminé d'après dosage par mesure d'absorbance à 280 nm. Les échantillons ont également été analysés par électrophorèse sur gel de polyacrylamide en présence de dodécylsulfate de sodium (SDS-PAGE), montrant dans chaque cas une bande intense à la taille attendue pour une Nanofitine monomérique témoignant d'une pureté satisfaisante (**Figure 32**).



II.2.3.3. Détermination des constantes d'affinité

Comme le suggérait le criblage en lysats bactériens, la fixation spécifique à la GFP par les Nanofitines purifiées a été confirmée par ELISA (**Figure 33A**), avec l'obtention d'un signal saturant ou de l'ordre du bruit de fond, respectivement en présence ou absence de GFP immobilisée. Cette fixation a également été caractérisée plus finement par interférométrie de couche biologique, avec différentes concentrations de GFP en solution après immobilisation de chaque Nanofitine sur les biocapteurs (**Figure 33B**), afin de calculer les constantes d'affinité de l'interaction. De façon intéressante, on peut noter que les huit Nanofitines purifiées présentent toutes des K_D nanomolaires (**Tableau 3**), avec les clones D8, C6 et C8 sous la barre des 2 nM grâce à des constantes cinétiques de dissociation (k_{off}) avantageuses (**Tableau 3, Figure 33B**).

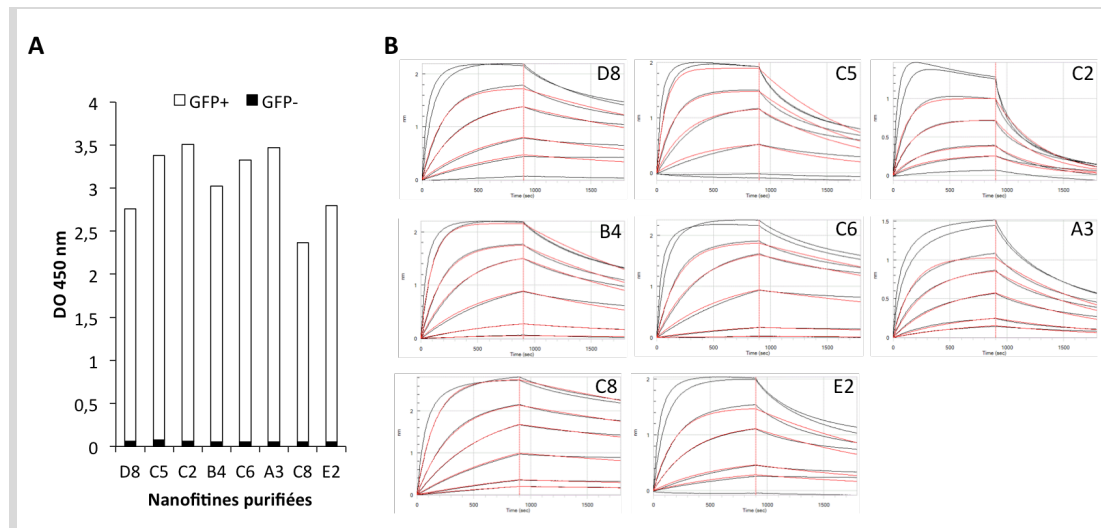


Figure 33: Profils de fixation des Nanofitines purifiées à la GFP, par ELISA et interférométrie. A) Confirmation de la fixation des Nanofitines purifiées à la GFP par ELISA en présence (barres blanches) ou absence (barres noires superposées) de StrepTagII-GFP immobilisée. B) Profils cinétiques d'association et de dissociation par interférométrie entre les Nanofitines purifiées, capturées sur les biocapteurs, et la GFP à différentes concentrations. Les mesures expérimentales sont tracées en noir, et les modèles ajustés sont tracés en rouge.

Tableau 3: Constantes d'affinité des Nanofitines anti-GFP purifiées et caractérisées.

	D8	C5	C2	B4	C6	A3	C8	E2
K_D ($\times 10^{-9}$ M)	1,82	4,68	9,31	3,12	1,61	4,9	1,83	3,25
k_{on} ($\times 10^5$ M ⁻¹ .s ⁻¹)	2,08	2,27	2,68	1,84	2,02	2,04	1,21	1,85
k_{off} ($\times 10^{-3}$ s ⁻¹)	0,38	1,06	2,49	0,57	0,33	1	0,22	0,6
R^2	0,9955	0,9846	0,991	0,9984	0,9978	0,9969	0,9995	0,9939

Les constantes d'affinité, mesurées par interférométrie, sont indiquées par les valeurs de K_D : constante de dissociation à l'équilibre; k_{on} : constante cinétique d'association; k_{off} : constante cinétique de dissociation; R^2 : coefficient de détermination du modèle ajusté.

II.2.3.4. Identification d'épitopes distincts sur la GFP

Les profils d'affinité des Nanofitines anti-GFP sont donc prometteurs, mais peu discriminants. Il est cependant intéressant de noter la dominance de certains acides aminés à des positions rendues aléatoires dans la banque employée lors de la sélection *in vitro*. Notamment, on peut observer les mutations K22L, W24T, V26F, T33H et T40W dans plus de la moitié des clones anti-GFP purifiés. Bien que deux de ces mutations n'aient pas été fortement enrichies entre le quatrième et sixième tour (de 26,0% à 22,7% pour K22L, de 32,3% à 36,4% pour V26F), les trois autres ont significativement augmenté leur représentation (de 46,9% à 61,4% pour T33H, de 28,1% à 68,2% pour W24T, et de 37,5% à 78,0% pour T40W). Du fait de ces différences de composition en acides aminés (**Figure 31**), nous avons formulé l'hypothèse que les Nanofitines partageant le moins d'identité (**Figure 34**) pouvaient potentiellement se fixer sur des régions différentes de la GFP, et plus particulièrement celles ne contenant pas les cinq mutations précédemment listées. Nous avons donc anticipé d'éventuels épitopes distincts pour C8, E2, ou les six autres Nanofitines.

Pour mettre en avant une divergence de région ciblée sur la surface de la GFP, les huit Nanofitines ont été utilisées dans une expérience de compétition en tandem mesurée par interférométrie (ou *epitope binning*; Torretta et al., 2012). Dans cette configuration, une Nanofitine est fixée jusqu'à saturation sur un biocapteur exposant de la GFP immobilisée à sa surface. Ainsi, les épitopes de la GFP ciblés par cette Nanofitine saturante se voient masqués par l'établissement de cette interaction, masquant aussi partiellement d'éventuels épitopes chevauchants. Ensuite, la mesure en temps réel du décalage de motif d'interférence permet d'identifier l'absence de compétition, donc la liaison à un épitope distinct, si une fixation de la

Nanofitine compétitrice est observée sur le complexe formé par la Nanofitine saturante liée à la GFP. Les résultats obtenus, dont un exemple représentatif est illustré **Figure 35A**, montrent que seule la Nanofitine C8 est apte à se fixer sur la GFP saturée par une autre Nanofitine.

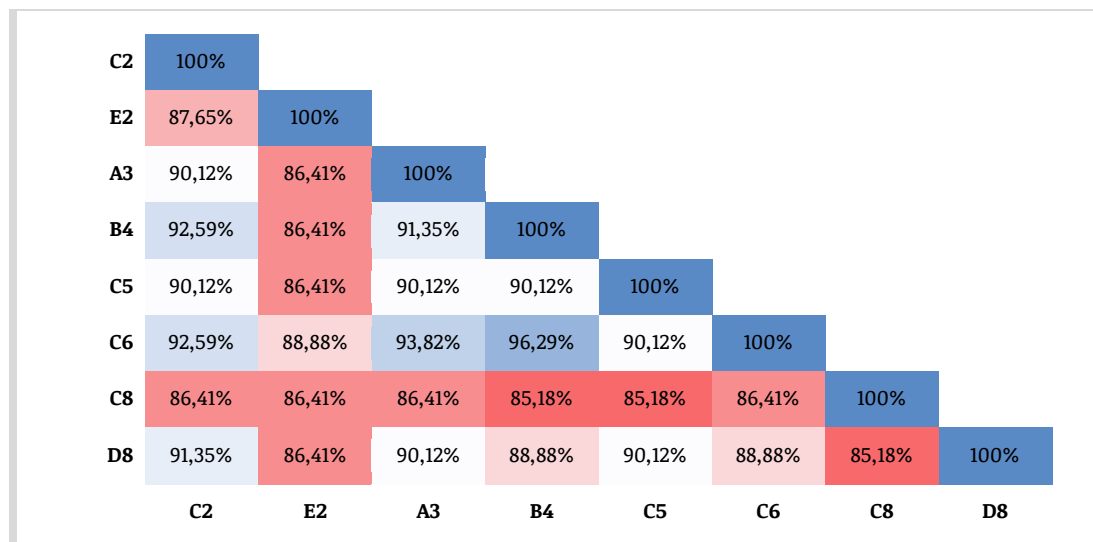


Figure 34: Matrice représentant les pourcentages d'identité entre les huit Nanofitines anti-GFP purifiées et caractérisées. Les pourcentages d'identité les plus élevés sont figurés en bleu, et les plus faibles en rouge.

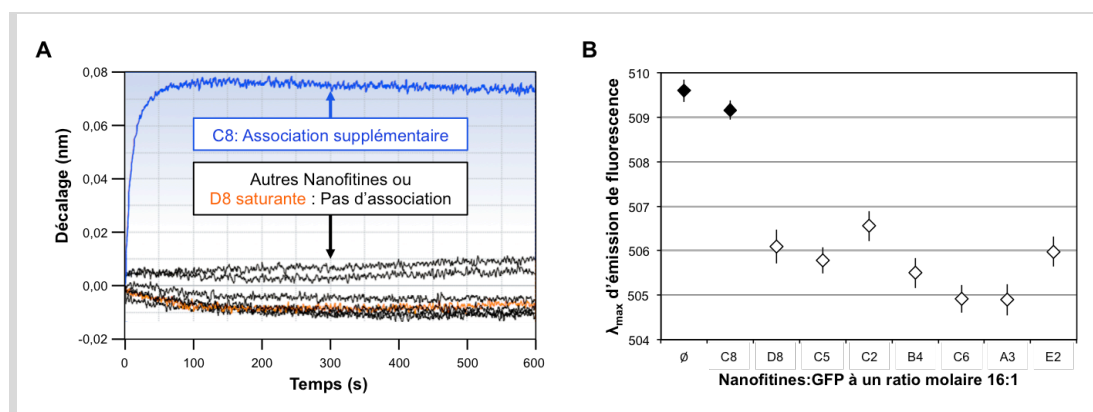


Figure 35: Mise en évidence d'épitopes distincts à la surface de la GFP. A) Exemple de résultat de compétition en tandem mesurée par interférométrie. Les courbes représentent la cinétique d'association des Nanofitines C2, E2, A3, B4, C5 et C6 (en noir), de la Nanofitine D8 (en orange), ou de la Nanofitine C8 (en bleu) sur un biocapteur avec GFP immobilisée préalablement saturé avec la Nanofitine D8. B) Longueurs d'onde d'émission maximales (λ_{max}) d'émission de fluorescence de la GFP en absence (\emptyset) ou présence de 16 excès molaires de Nanofitine purifiée. Les conditions ne modifiant pas la fluorescence de la GFP sont figurées en noir ($509 \text{ nm} \leq \lambda_{max} \leq 510 \text{ nm}$), et celles induisant une diminution de la longueur d'onde maximale de fluorescence sont figurées en blanc ($\lambda_{max} \leq 507 \text{ nm}$).

En parallèle de ces expériences de compétition, la recherche de molécules de fixation à la GFP avec des épitopes connus a permis d'identifier l'existence de fragments d'anticorps à chaînes lourdes, appelés Nanobodies, capables de moduler les propriétés de fluorescence de la GFP (Kirchhofer et al., 2010; Kubala et al., 2010). En plus de se lier à leur cible, deux des Nanobodies décrits induisent un décalage d'absorbance de la GFP, ce qui permet à l'un d'en augmenter

l'émission de fluorescence et à l'autre de la diminuer. Connaissant la possibilité de moduler les propriétés de fluorescence de la GFP par l'interaction avec une protéine d'affinité, les spectres d'émission de fluorescence de la GFP ont été acquis en présence de 16 équivalents molaires de Nanofitine. Curieusement, la distinction entre C8 et les sept autres Nanofitines anti-GFP a également été observée lors des mesures. En effet, la longueur d'onde maximale d'émission de fluorescence de la GFP est comprise entre 509 et 510 nm en absence de Nanofitine ou en présence de C8, alors qu'elle est décalée à moins de 507 nm en présence des autres Nanofitines (**Figure 35B**). Ces observations suggèrent donc que la Nanofitine anti-GFP C8 est dirigée contre un épitope distinct de celui fixé par les autres Nanofitines sélectionnées.

II.2.3.5. Sites de fixation à la GFP choisis pour le transfert de charpente

Grâce à la sélection *in vitro* de variants Nanofitines spécifiques de la GFP, nous avons donc pu identifier des molécules d'affinités nanomolaires, dont le site de fixation a été généré à partir de l'introduction de mutations aléatoires dans le feuillet beta de Sac7d. Après caractérisation de ces clones, les Nanofitines C6 et C8 ont été désignées comme candidats au transfert de charpente pour en maximiser les chances de réussite, en anticipation d'une éventuelle diminution d'affinité mais également pour explorer différentes régions de la GFP lors du processus. De fait, ces Nanofitines sont celles qui possèdent les meilleures affinités envers la GFP, présentant notamment les k_{off} les plus avantageux, tout en ciblant deux épitopes distincts à la surface de la GFP.

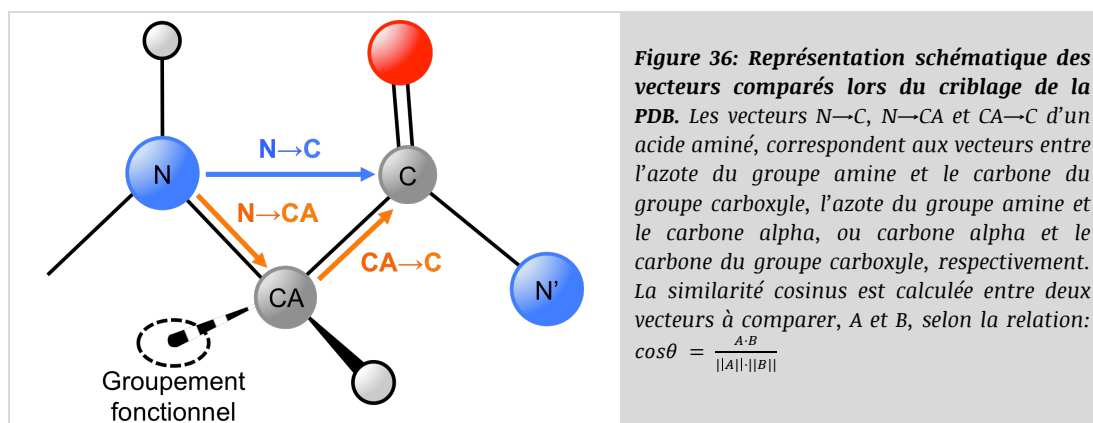
Dans le présent manuscrit, les régions comprenant les résidus 21-26, 29-33 et 40-46 des Nanofitines anti-GFP seront nommées "feuillets anti-GFP".

II.3. Charpentes protéiques humaines hôtes de feuillet anti-GFP

II.3.1. Recherche structurale de feuillets hôtes

La charpente protéique des Nanofitines a été prouvée très robuste malgré l'introduction de mutations dans le second feuillet beta de Sac7d, démontrant notamment une résistance à des températures élevées (Mouratou *et al.*, 2007) jusqu'à autoriser leur stérilisation par autoclavage à 121°C (Huet *et al.*, 2015). Toutes les protéines ne présentent pas un profil aussi résistant, mais la

majorité reste cependant sujette à la corrélation entre labilité locale et structures secondaires, qui établit que les structures beta, mieux organisées, sont généralement plus rigides que les structures alpha ou désordonnées (Peticaroli et al., 2013, 2014). Cet aspect a été mis au centre de la conception de notre démarche d'humanisation des Nanofitines, puisque la rigidité de l'agencement spatial des acides aminés fonctionnels présentés sur un feuillet beta devrait les rendre plus facilement identifiables. D'une part, ceci nous permettrait de ne pas dévier de façon importante des structures de Sac7d déjà résolues. D'autre part, ceci devrait augmenter la précision atteinte lors de la recherche de charpentes hôtes, car celles-ci sont représentées dans une conformation relativement figée dans leurs structures cristallographiques. Par conséquent, cette observation nous a incité à transférer uniquement les 11 résidus mutés des feuillets anti-GFP des Nanofitines C6 et C8 décrites précédemment.



Pour identifier les protéines aptes à recevoir ces domaines anti-GFP, un criblage de la banque de données sur les protéines du *Research Collaboratory for Structural Bioinformatics (Protein Data Bank, ou PDB)* a été réalisé par Yves-Henri Sanejouand, consistant en une recherche itérative de vecteurs $N \rightarrow CA$ et $CA \rightarrow C$ (**Figure 36**) proches de ceux des feuillets anti-GFP générés sur Sac7d (identifiant PDB: 1AZP; Robinson et al., 1998). En bref, les coordonnées spatiales du second feuillet beta de Sac7d ont été recherchées à partir des atomes de la chaîne principale (azote du groupe amine, carbone alpha et carbone du groupe carboxyle, respectivement désigné N, CA et C) des 11 résidus mutés (exposés en surface) ainsi que des 7 résidus enfouis (résidus V23, R25, V30, F32, G41, G43 et V45). La différence de géométrie a été estimée par calcul du RMS entre ces atomes, avec les RMS de plus faibles valeurs indiquant des géométries de feuillet semblables entre les charpentes

donneuse et hôte. Des indications supplémentaires sur les similarités d'environnement des feuillettes identifiées lors du criblage de la PDB ont également été calculées. Notamment, la différence d'accessibilité induite par le transfert a été définie en calculant la valeur moyenne des surfaces accessibles par résidu, puis en déterminant le RMS de ces valeurs, devant indiquer les accessibilités voisines de celle du feuillet de Sac7d via les valeurs les plus faibles. La différence d'orientation des résidus a également été estimée par calcul de la similarité cosinus pour l'ensemble du feuillet beta ou uniquement le résidu le plus ré-orienté (cosinus le plus proche de 0 par rapport au feuillet recherché). Le calcul du cosinus de l'angle entre les directions des deux vecteurs permet d'identifier les cas de similarité maximale (correspondant à un angle de 0, soit un cosinus prenant la valeur de 1). Enfin, des critères additionnels ont été intégrés à ce criblage afin d'isoler des charpentes protéiques possédant un profil favorable à leur utilisation *in vivo*, étant donc potentiellement compatibles avec un usage thérapeutique, tout en autorisant une production dans l'infrastructure en place au laboratoire:

- 1) Protéines humaines: pour proposer une alternative à l'origine archéobactérienne des Nanofitines, tout en étant exempte de propriété intellectuelle dans la mesure du possible.
- 2) Protéines annotées comme monomériques: pour limiter le risque de masquage du site de transfert des feuillettes anti-GFP, et pour bénéficier de protéines monovalentes et plus homogènes lors de leur production.
- 3) Protéines sans ponts disulfures: pour limiter la sensibilité aux conditions réductrices et pour permettre une production simple en *E. coli*.
- 4) Structures avec une résolution de 2 Å ou moins: pour écarter les structures les moins fiables.
- 5) Protéines avec maximum 90% d'identité de séquence: pour réduire le biais de sur-représentation des protéines les plus facilement cristallisables ou les plus étudiées (*Gerstein, 1998; Peng et al., 2004; Xie et Bourns, 2005; Le Gall et al., 2007; Cossio et al., 2010*).
- 6) RMS de 1 Å ou moins entre le feuillet à remplacer et celui à transférer: pour identifier les feuillettes dans lesquels le squelette carboné des feuillettes anti-GFP serait le moins modifié.

Après criblage, 203 structures de protéines humaines ont été identifiées comme remplissant l'ensemble de ces critères.

II.3.2. Comparaison avec les structures en OB-fold

Afin d'examiner la population des 203 structures de protéines humaines identifiées, celle-ci a été comparée à l'ensemble de structures partageant un motif de repliement commun avec Sac7d disponibles dans la PDB, humaines ou non (structures de type OB-fold avec une résolution maximale de 2 Å, et une seule chaîne ou unité asymétrique). Les deux groupes obtenus par criblage des vecteurs N→CA et CA→C ont finalement comporté 203 et 103 membres, respectivement, pour les protéines humaines et les protéines de type OB-fold.

Curieusement, nous pouvons observer une forte hétérogénéité dans les feuillets beta issus des protéines en OB-fold, générant des RMS calculés à partir du feuillet beta anti-GFP des Nanofitines majoritairement au dessus du seuil de coupure de 1 Å imposé pour les protéines humaines (0,79 à 2,12 Å, avec $1,47 \pm 0,28$ Å de RMS moyen). En plus de démontrer la diversité et la flexibilité observables dans les structures en OB-fold, cette répartition des RMS (**Figure 37** "RMS") pourrait être considérée comme un bon indicateur de la similarité structurale des feuillets de Sac7d et de ceux découverts dans les protéines humaines criblées.

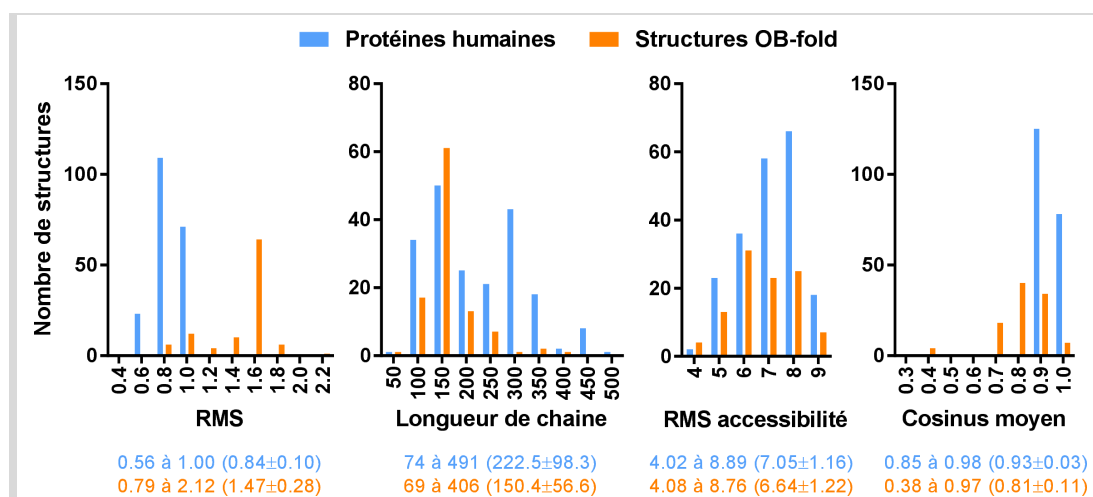


Figure 37: Comparaison entre structures humaines criblées et structures en OB-fold. Distributions des scores obtenus lors du criblage de la PDB sur les 203 structures de protéines humaines (bleu) ou les 103 structures en OB-fold (orange). Les valeurs minimales et maximales de chaque groupe sont indiquées sous les histogrammes, avec leur moyenne et écart-type entre parenthèses.

Cependant, nous constatons que les deux populations confrontées ne concernent pas des protéines de mêmes tailles. A l'instar des Nanofitines, les protéines de type OB-fold sont en règle

générale plus petites que les domaines résolus des protéines humaines étudiées (**Figure 37** “Longueur de chaîne”), avec en moyenne 150 acides aminés (± 57) contre 223 (± 98). Cette différence de taille, impliquant des structures potentiellement plus complexes, pourrait être la cause de la légère diminution de l’accessibilité des feuillets par rapport à leur environnement sur Sac7d (**Figure 37** “RMS accessibilité”), estimée par une augmentation des RMS d’accessibilité calculés pour les protéines humaines ($7,05 \pm 1,16 \text{ \AA}^2$ en moyenne) vis-à-vis de ceux obtenus avec les protéines de type *OB-fold* ($6,64 \pm 1,22 \text{ \AA}^2$ en moyenne). Malgré tout, les structures humaines criblées présentent une meilleure similarité cosinus (**Figure 37** “Cosinus moyen”) que l’ensemble des *OB-fold* étudiées, ce qui devrait indiquer des feuillets avec une orientation globale plus proche de celle des Nanofitines.

Cette comparaison illustre de légères différences entre des protéines filtrées selon leur origine (domaines protéiques humains) ou leur repliement (protéines de type *OB-fold*). Néanmoins, nous pouvons constater que des scores avantageux peuvent être identifiés dans les deux cas, indépendamment du groupe étudié. Pour conserver ce point de comparaison à l’échelle expérimentale, nous avons inclus comme référence la charpente de Sso7d (identifiant PDB: 1C8C; *Su et al., 2000*), une protéine archéobactérienne homologue de Sac7d générant un RMS de $0,178 \text{ \AA}$ et un RMS d’accessibilité de $2,44 \text{ \AA}^2$.

II.3.3. Sélection des charpentes hôtes

A l’issue du criblage des protéines humaines de la PDB, nous avons pu constater une présence importante de domaines redondants, principalement représentés par le dixième domaine de la fibronectine de type III (10Fn3) ou des protéines de la famille des lipocalines. Malheureusement, ces charpentes sont l’objet de brevets pour la génération de protéines d’affinité, respectivement connues sous les dénominations adnectines ou anticalines. De ce fait, ces protéines ne satisfaisaient pas la condition d’absence de propriété intellectuelle intégrée au cahier des charges des travaux de thèse. Malgré tout, nous avons conservé une lipocaline (identifiant PDB: 1LPJ; *Folli et al., 2002*) sélectionnée parmi les protéines aboutissant aux meilleurs scores de RMS.

Trois protéines humaines ont ainsi été retenues (identifiants PDB: 1LPJ, 4F7H (Yan Liu et al., 2012), 1QNT (Wibley et al., 2000)), et leurs scores ont été comparés à ceux calculés en cas de transfert de Sac7d (identifiant PDB: 1AZP) vers Sso7d (identifiant PDB: 1C8C) ou vers la structure d'une Nanofitine déjà résolue (identifiant PDB: 2XIW; Béhar et al., 2013), tel que listé dans le **Tableau 4**.

Tableau 4: Meilleurs scores du criblage dans la PDB.

Protéine	PDB	Acides aminés			RMS (Å)	Cos	Cos min	RMS Accessibilité (Å ²)	Ordre des brins	Commentaire
		Totaux	Trouvés	Mutés						
Sac7d	1AZP	66	18	0	0,000	1,00	1,00	0,00	=	Feuillet recherché
Sso7d	1C8C	64	18	1	0,178	1,00	0,98	2,44	=	Contrôles homologues
Mutant de Sac7d	2XIW	64	18	11	0,356	0,99	0,95	4,69	=	
Protéine de liaison de rétinolide 7 (Lipocaline)	1LPJ	133	18	17	0,559	0,96	0,57	5,34	=	Meilleurs résultats
Protéine homologue de la famille fermitine 2	4F7H	135	18	17	0,627	0,96	0,74	5,36	=	
Protéine de réparation de l'ADN	1QNT	166	18	17	0,630	0,97	0,89	5,29	≠	

PDB: identifiant PDB. Les acides aminés totaux, trouvés et mutés représentent, respectivement, le nombre de résidus totaux de la structure cristallisée, le nombre de résidus trouvés correspondant au feuillet beta de Sac7d (max: 18), et le nombre de résidus mutés dans le cas d'un transfert de brin complet (résidus exposés et enfouis). RMS: Déviation de la racine de la moyenne des carrés des positions atomiques, en Å. Cos: cosinus moyen entre les vecteurs N→C de la structure et ceux de 1AZP. Cos min: valeur minimale de cosinus par couple de vecteurs N→C. RMS Accessibilité: Déviation de la racine de la moyenne des carrés d'accessibilité, en Å². Ordre des brins: indique si les brins beta sont dans le même ordre (=), ou non (≠), dans le feuillet identifié.

Malgré des RMS jusqu'à 3,5 fois plus importants vis-à-vis du transfert de 1AZP vers 1C8C, ce score représente moins du double de celui généré par transfert de 1AZP vers 2XIW. De plus, l'orientation globale des acides aminés, traduite par des vecteurs similaires, semble propice au transfert puisque les résultats de similarité cosinus sont proches de 1, malgré certains résidus orientés très différemment (identifiés dans la colonne "Cos min"). Sur la base de ce critère d'orientation, la structure 1QNT semblerait d'un intérêt particulier, alors que 1LPJ présente un potentiel de déstabilisation plus important. De même, les RMS d'accessibilité calculés pour les 3 structures humaines proposées (5,3 Å en moyenne) sont parmi les plus intéressants de ceux obtenus, se situant sous la moyenne observée à 7,1 Å pour les protéines criblées. Ceci représente des

potentiels d'accessibilité parmi les plus favorables, comme également constaté visuellement sur les représentations des brins beta identifiés sur les structures des protéines sélectionnées pour le transfert (**Figure 38**). Pour finir, nous avons anticipé une expression efficace en système bactérien puisque ces 3 structures sont de taille modeste, comprises entre 133 et 166 acides aminés, et ont pu être exprimées en *E. coli* en vue de leur cristallisation (Folli *et al.*, 2002; S. Liu *et al.*, 2007; Wibley *et al.*, 2000).

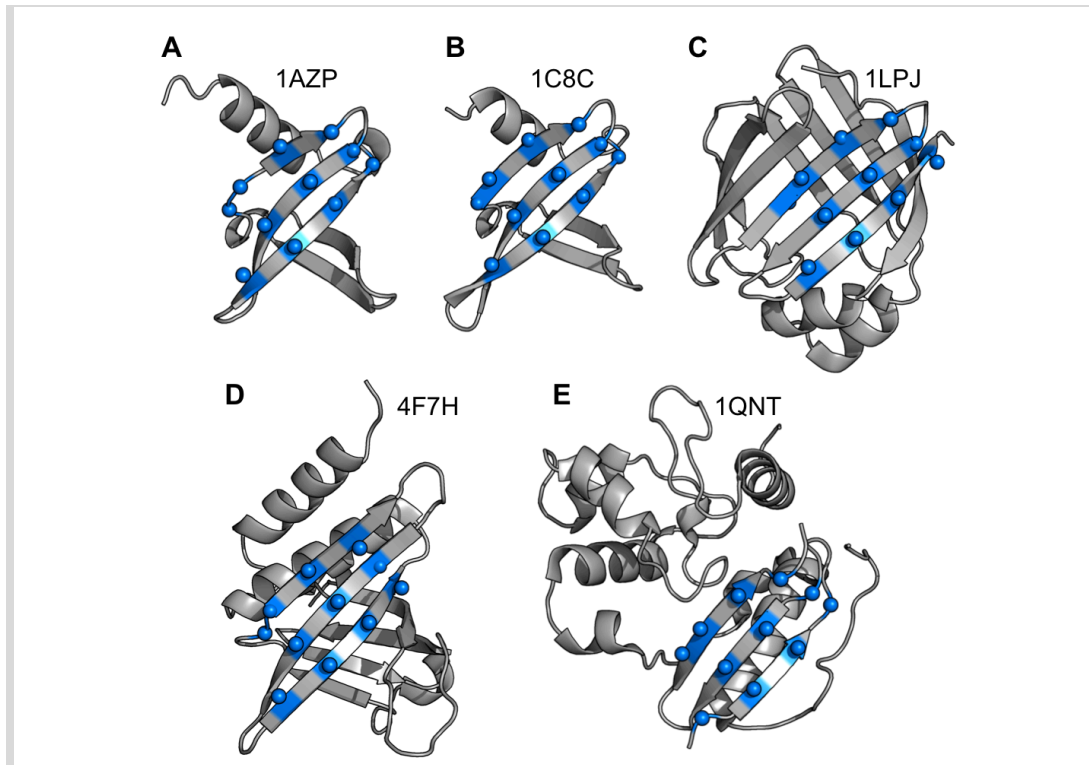


Figure 38: Structure des charpentes hôtes des feuillets anti-GFP. Représentations schématiques des structures tridimensionnelles de Sac7d (A), Sso7d (B) et des trois charpentes humaines retenues pour le transfert de feuillet anti-GFP (C, D, E). Les résidus à muter, correspondant aux acides aminés de surface des feuillets anti-GFP, sont représentés en bleu et leur carbone alpha par une sphère.

II.4. Expression sous forme soluble des charpentes humaines

II.4.1. Obtention et clonage des séquences codantes

Les séquences codant pour les protéines humaines, synthétisées par un laboratoire tiers, ont été sous-clonées par digestion via les enzymes de restriction BamHI et HindIII puis par ligation dans un vecteur d'expression standard, nommé pAFG01 (**Figure 39**). Après séquençage, un plasmide a été validé pour chaque construction (**Figure 40**), permettant l'expression de chaque protéine recombinante en *E. coli* ainsi que sa purification grâce à une étiquette poly-histidine à son extrémité N-terminale.

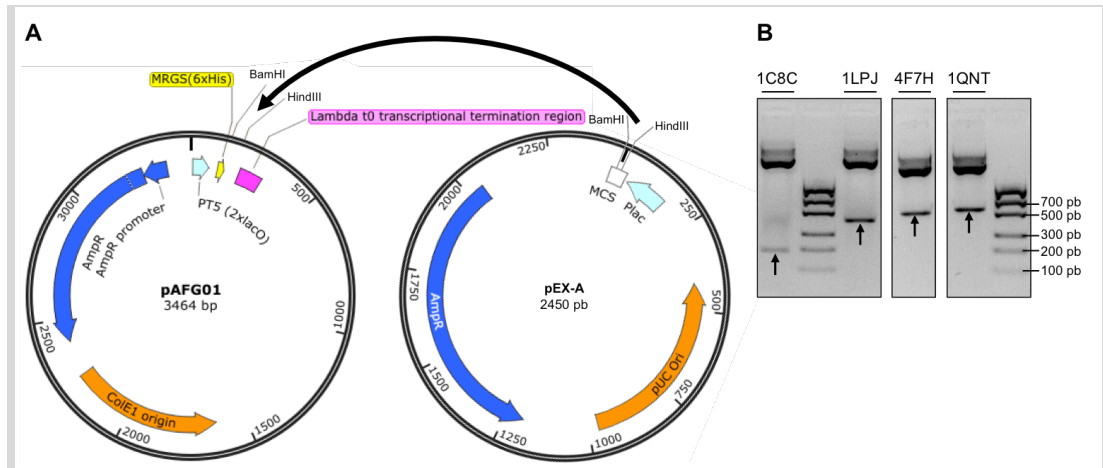


Figure 39: Sous-clonage des gènes codant pour les protéines humaines hôtes. A) Cartes du vecteur d'expression pAFG01 et du plasmide pEX-A. Les séquences codantes des protéines hôtes ont été sous-clonées dans pAFG01 à partir pEX-A, par digestion puis ligation entre les sites de restriction BamHI et HindIII. B) Exemple de profils des produits de digestions par BamHI et HindIII des plasmides pEX-A contenant les séquences codantes de mutants de 1C8C, 1LPJ, 4F7H et 1QNT après électrophorèse sur gels d'agarose. Les flèches indiquent les inserts digérés de taille attendue.

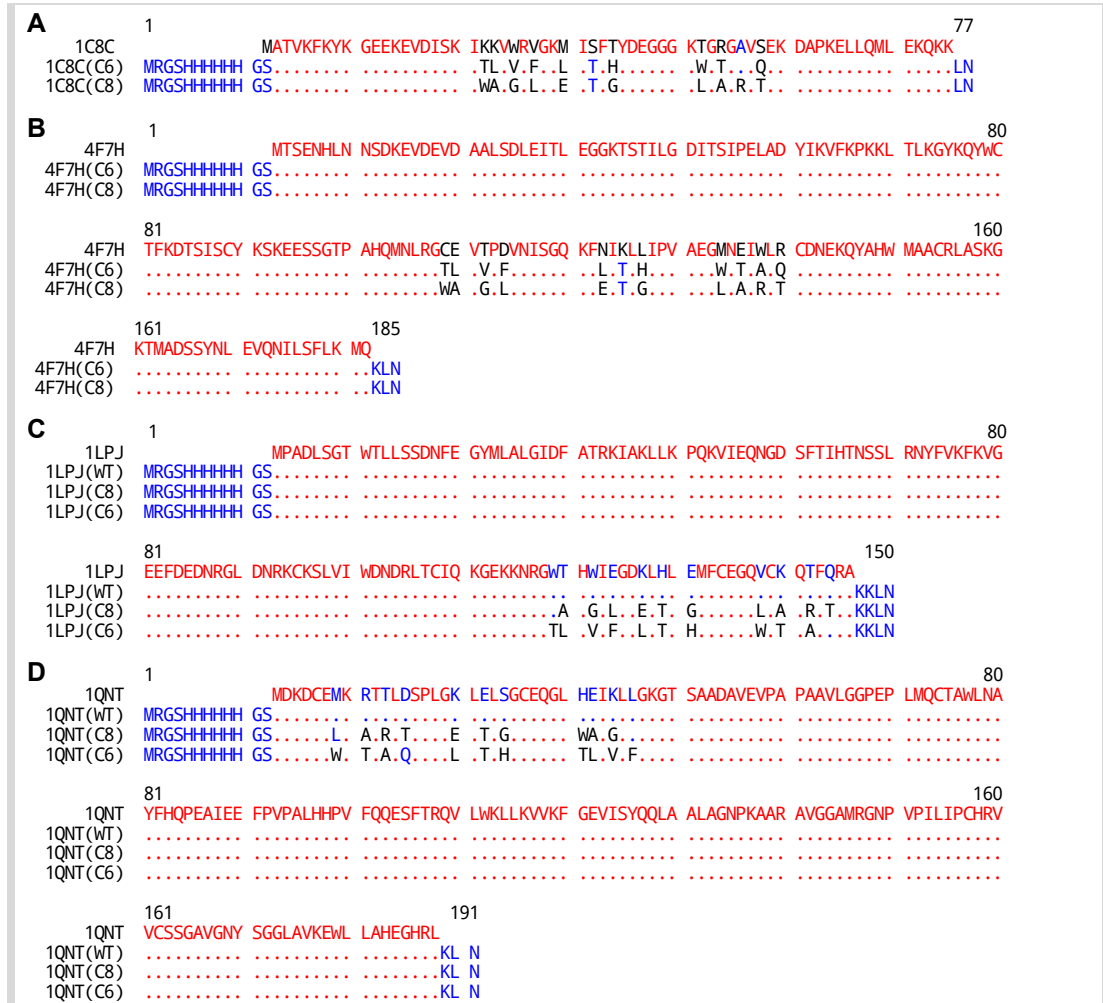


Figure 40: Alignements de séquence des protéines hôtes. Séquences des variants de 1C8C (A), 4F7H (B), 1LPJ (C) et 1QNT (D) clonés et exprimés en *E. coli*, alignées aux séquences de référence des charpentés. WT: Protéines sauvages, clonées sans mutation en dehors de leurs extrémités N- et C-terminales. C6: Variants ayant été greffés du feuillet anti-GFP de C6. C8: Variants ayant été greffés du feuillet anti-GFP de C8.

II.4.2. Expression en bactérie et impact du transfert sur la solubilité

II.4.2.1. Cultures standards

Conformément à nos attentes, il a été possible d'obtenir les variants Sso7d contenant les feuillets anti-GFP des Nanofitines C6 et C8 sous forme soluble, d'après l'analyse SDS-PAGE réalisée (**Figure 41**). Une bande intense est en effet présente dans la fraction purifiée par IMAC (fraction EL), à la taille attendue de ~9 kDa.

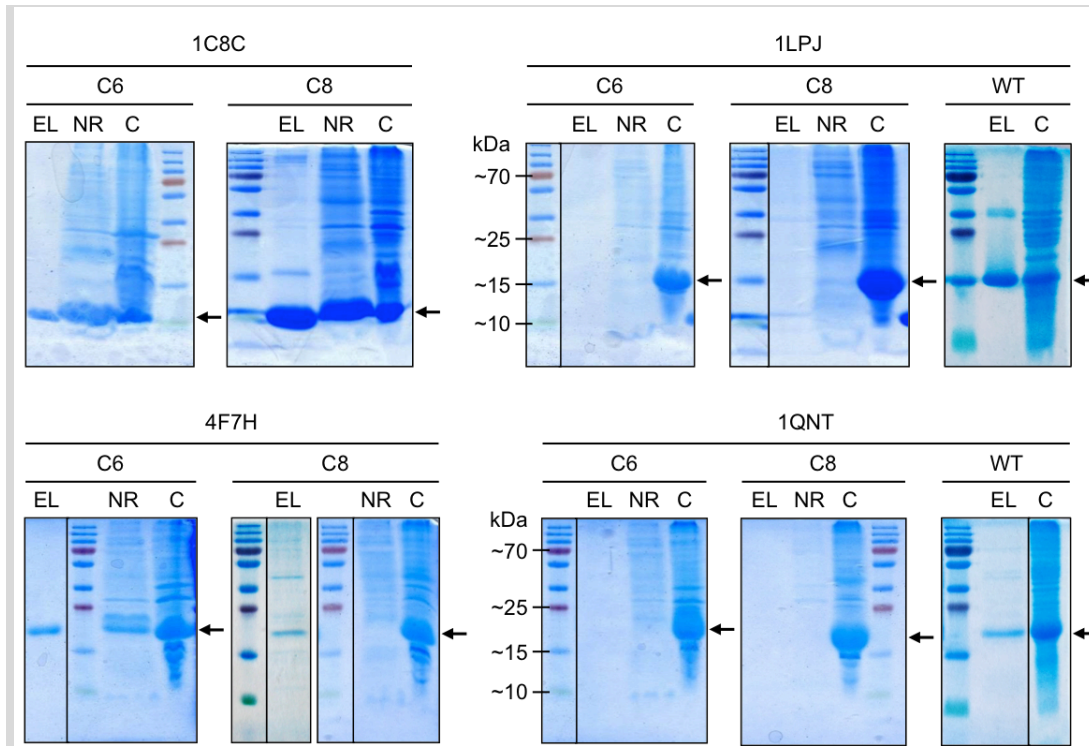


Figure 41: Profils SDS-PAGE avec 2 µg de protéine purifiée par IMAC (EL) pour chaque variant exprimé en *E. coli*, après coloration au bleu de Coomassie. Des dépôts de 16 µL ont été réalisés pour les fractions non retenues lors de la chromatographie d'affinité (NR) ou les fractions EL trop diluées pour obtenir 2 µg. Des dépôts ont également été réalisés à partir d'échantillons de culot de lyse (C). La distance de migration attendue pour chaque charpente monomérique est indiquée par une flèche.

Dans les mêmes conditions standards d'expression de Nanofitines (*E. coli* DH5α LacI^q inoculées dans 200 mL de milieu 2YT à 37°C, induites en phase exponentielle de croissance pendant la nuit à 30°C par ajout d'IPTG), un profil similaire a été obtenu pour les transferts sur 4F7H (~26 kDa), tandis qu'aucune protéine soluble n'a pu être détectée à la taille attendue lors du transfert sur 1LPJ ou 1QNT (respectivement, ~20 et ~27 kDa). Bien qu'aucune bande aux tailles attendues n'ait été observée dans la fraction éluee ou non-retenue par IMAC (EL et NR, respectivement), une bande d'intensité plus élevée a été observée dans la fraction non soluble du lysat bactérien (pistes

C), suggérant une insolubilité de ces protéines recombinantes après transfert d'un feuillet anti-GFP. La responsabilité du transfert a d'ailleurs été confirmée par la production des protéines 1LPJ et 1QNT sans mutation (WT), permettant l'obtention de protéines solubles purifiées par IMAC dans les mêmes conditions.

II.4.2.2. Exploration de conditions augmentant la solubilité

Face à l'insolubilité des charpentes de 1LPJ et 1QNT ayant été greffées d'un feuillet anti-GFP issu d'une Nanofitine, différentes stratégies ont été explorées en vue de résoudre ce problème, tel que résumé sur la **Figure 42**.

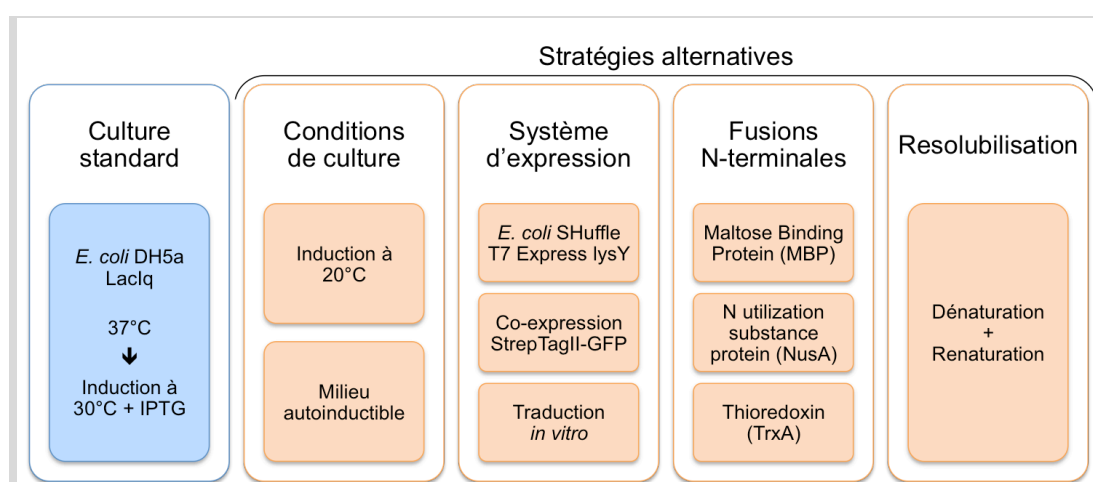


Figure 42: Diagramme synthétique des conditions de culture standard et des stratégies alternatives réalisées pour contourner l'insolubilité des variants de 1LPJ et 1QNT.

Des approches évitant la modification de la protéine recombinante ont tout d'abord été abordées. En modifiant les conditions de culture, l'expression des protéines recombinantes a été réalisée à température réduite lors de la période d'induction (20°C au lieu de 30°C), ce qui a été décrit comme une piste pouvant limiter l'agrégation *in vivo* (Schein, 1989). La composition du milieu a également été modifiée dans une autre condition de culture, en utilisant un milieu auto-inductible devant retarder l'expression des protéines d'intérêt dans les premières phases de croissance bactérienne, et donc limiter le détournement des ressources cellulaires, par induction de la production de lactose perméase après consommation du glucose en plusieurs heures (Studier, 2005).

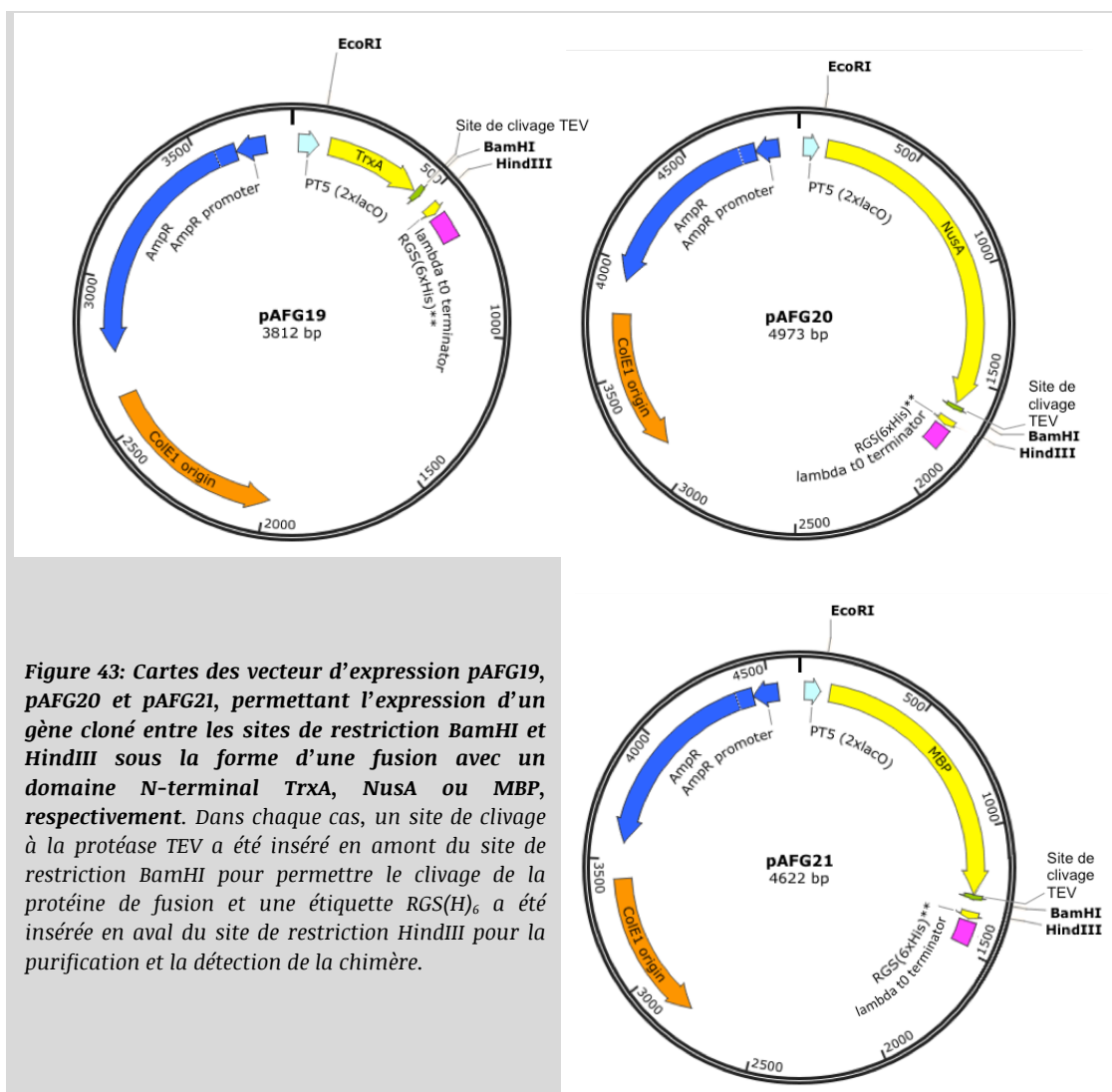
L'ingénierie des protéines à solubiliser a aussi été contournée par l'utilisation de systèmes d'expression alternatifs. La souche *E. coli* DH5a LacI^q a été substituée par la souche *E. coli* SHuffle

T7 Express lysY, enrichie en protéines chaperonnes DsbC (Levy et al., 2001) ainsi que déficiente en protéases lon (Gottesman et al., 1981) et protéases membranaires ompT (Grodberg et Dunn, 1988), développée pour produire des protéines humaines dans le compartiment cytoplasmique (y compris celles nécessitant d'établir des ponts disulfures). De même, un système de co-expression avec le partenaire d'interaction attendu, la StrepTagII-GFP, a été mis en place pour remédier à l'exposition de résidus hydrophobes à la surface des protéines et pour réduire le risque d'insolubilité (Kurochkina et Mesyanzhinov, 1999). Enfin, une approche de traduction *in vitro* a également été explorée, à partir d'extrait de S30 couramment utilisé en *ribosome display* (Amstutz et al., 2006), pour tenter des micro-productions en solution moins complexe en vue de la caractérisation rapide des protéines exprimées sans nécessiter de purification.

Des stratégies additionnelles nécessitant l'ingénierie des protéines à exprimer ont été mises en œuvre, via la fusion de domaines protéiques décrits comme ayant un impact positif sur l'expression, la stabilité et/ou la solubilité de protéines chimériques. Les gènes des protéines à solubiliser ont été exprimés après leur sous-clonage dans des vecteurs d'expression dérivés de pAFG01, ajoutant un domaine TrxA (Kapust et Waugh, 1999), NusA (Davis et al., 1999) ou MBP (Jacquet et al., 1999) à l'extrémité N-terminale de la protéine d'intérêt (**Figure 43**).

En dernier lieu, une approche de resolubilisation des protéines insolubles a été appliquée, d'une façon similaire aux protocoles de purification à partir de corps d'inclusion. Après dénaturation via l'utilisation d'urée 8 M comme agent chaotrope, une renaturation a été tentée par application d'un gradient décroissant d'urée sur les protéines immobilisées sur résine de nickel (Lemercier et al., 2003) ou par dilution rapide des protéines en solution (Yokoyama et al., 2002). Dans le premier cas, l'agent de solubilisation est graduellement retiré par chromatographie alors que la protéine d'intérêt est immobilisée, ce qui aboutit couramment à de forts rendements de protéines actives, y compris avec des concentrations de l'ordre du mg/ml. De plus, la chromatographie en conditions dénaturantes permet d'éliminer les protéines contaminantes avant renaturation. Dans le second cas, la dilution rapide de l'agent chaotrope réduit le temps d'exposition à des concentrations intermédiaires auxquelles les protéines à renaturer peuvent

être dans un état non replié sans pour autant être totalement dénaturées. De tels états intermédiaires de repliement rendent les protéines plus susceptibles à l'agrégation, phénomène limité par les méthodes de dilution rapide.



Malheureusement, le bilan de l'ensemble de ces stratégies est sans appel: les charpentes de 1QNT et ILPJ sont surexprimées et accumulées uniquement sous forme insoluble malgré les diverses pistes explorées. Toutes ces conditions, après analyse par SDS-PAGE, aboutissent à des profils semblables à ceux obtenus en conditions standards, avec pour seules protéines solubles des protéines endogènes du système d'expression. Ces résultats seront à nouveau discutés de manière plus extensive dans la section "II.5. Conclusions et observations sur la méthodologie de transfert de feuillets", accompagnés d'une analyse bioinformatique visant à expliquer et anticiper ce phénomène de surexpression non soluble.

II.4.3. Impact du transfert sur les structures 1C8C et 4F7H

Les protéines solubles et exposant un feuillet anti-GFP à leur surface qui ont été recueillies possèdent donc les charpentes de Sac7d (1AZP), de son homologue Sso7d (1C8C) ou d'un domaine de type fermitine d'une protéine humaine (4F7H). La comparaison des cinétiques de capture sur nickel et d'association à la GFP mesurées par interférométrie (dont un exemple est illustré pour le feuillet de C6, **Figure 44**) confirme dans un premier temps la capture efficace de la Nanofitine contrôle, des variants Sso7d, et des variants de la structure 4F7H (étape 1). Cette étape corrobore également l'absence de protéine soluble à capturer lors de l'utilisation des charpentes 1QNT et 1LPJ, générant des signaux de l'ordre de celui observé avec le biocapteur de référence. Une fois capturées, les protéines recombinantes ont été exposées à une solution de GFP pour valider leur capacité de liaison (étape 3). Bien qu'une fixation soit effectivement observée pour les variants de Sso7d avec les feuillets anti-GFP de C6 ou de C8, aucun signal d'association n'est détecté après capture des mutants de la structure 4F7H, aboutissant à des signaux similaires à ceux de la Nanofitine non-significative ou du biocapteur de référence. Cette absence de fixation a été confirmée par ELISA avec des concentrations jusqu'à de 1,6 nM à 5 μ M des mutants en solution, validant la perte d'affinité lors du transfert de Sac7d vers 4F7H.

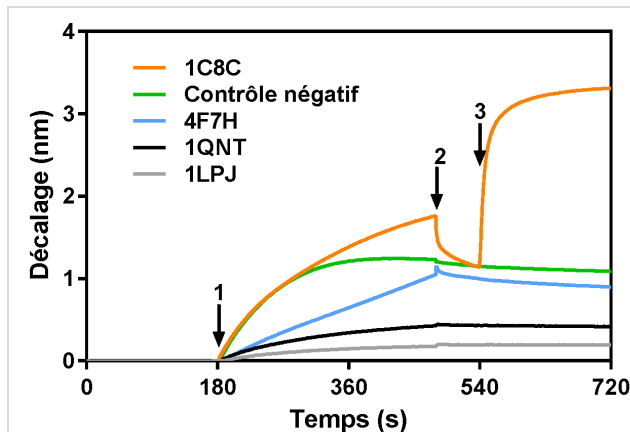


Figure 44: Profils cinétiques mesurés par interférométrie de couche biologique après traduction in vitro des variants greffés du feuillet anti-GFP de C6.
 1) Capture des produits de traduction sur les biocapteurs Ni-NTA, par leur étiquette polyhistidine. 2) Ligne de base avant association. 3) Interaction avec la GFP en solution. Contrôle négatif: produit de traduction d'une Nanofitine non spécifique.

Malgré l'impossibilité de solubiliser ou fixer les protéines humaines mutées à la GFP, les variants Sso7d greffés du feuillet anti-GFP de C6 ou C8 ont montré une conservation de leur capacité de fixation à la GFP, comme confirmé par ELISA. De façon intéressante, cette interaction s'est traduite par une co-purification du variant Sso7d et de la GFP par chromatographie d'affinité dans les conditions de co-expression, que ce soit par l'intermédiaire de leur étiquette

polyhistidine ou StrepTagII, respectivement. La caractérisation plus fine du transfert du feuillet anti-GFP de C8, effectuée par interférométrie (**Figure 45**), a cependant mis en avant une diminution par 6 du K_D , essentiellement du fait de l'augmentation de la constante cinétique de dissociation (d'un facteur 10). Cette perte d'affinité, bien que modeste, souligne la finesse engagée dans le transfert de domaines de fixation entre deux charpentes protéiques.

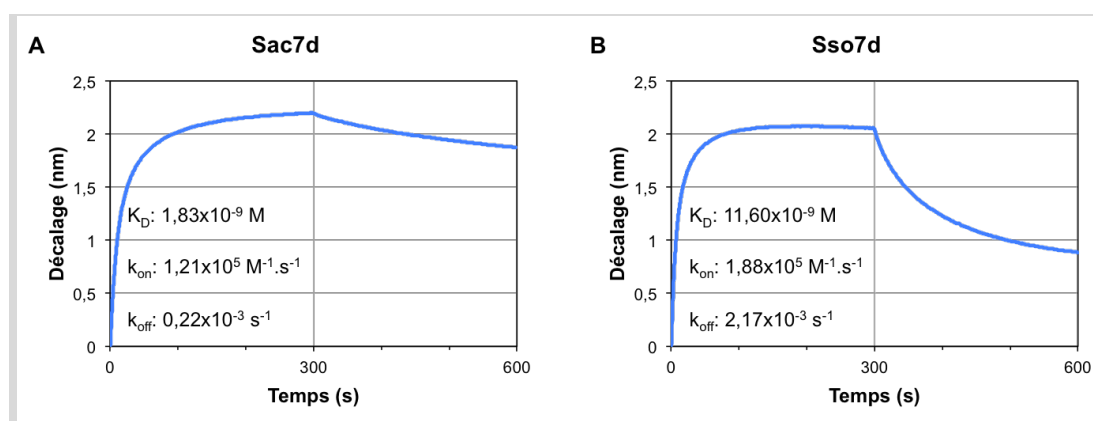


Figure 45: Comparaison des profils cinétiques d'association et dissociation du feuillet anti-GFP C8 sur les charpentes A) Sac7d et B) Sso7d, mesurés par interférométrie. Profils mesurés après capture des variants de Sac7d ou Sso7d sur un biocapteur Ni-NTA, avec 500 nM de GFP en solution. Les valeurs de K_D , k_{on} et k_{off} ont été calculées en présence de gamme de concentration de GFP de 100 nM à 1,56 nM.

Malgré une forte proximité de séquence et de structure, les charpentes de Sac7d et Sso7d ont donc un impact non négligeable sur le domaine de fixation qu'elles présentent à leur surface. Ce phénomène a par ailleurs été décrit par Pecorari et ses collaborateurs (Béhar *et al.*, 2014) lors de la stabilisation, vis-à-vis du pH et de la température, d'une Nanofitine ciblant les immunoglobulines de type G (IgG) par transfert de la charpente de Sac7d à celle de Sso7d, plus thermostable à l'état natif. Cet exemple de transfert, effectué sur une portion du même feuillet beta ainsi que de deux boucles, a résulté en une augmentation de K_D de 34 nM à 1,5 μM , témoignant de la sensibilité d'une telle approche. Cette diminution d'affinité confirme également l'intérêt de pouvoir bénéficier de charpentes dont les variants puissent être directement générés par un procédé de sélection, comme publié par Rao et ses collaborateurs par exemple pour Sso7d (Gera *et al.*, 2011), pour leur conférer des propriétés attendues dès leur génération et sans risquer de les dégrader. De manière générale, la confrontation à des pertes d'affinité est une problématique récurrente lors de l'ingénierie de protéines d'affinité, comme illustré par de nombreux exemples de transfert de boucles entre charpentes, ou lors de l'humanisation

d'anticorps (comme préalablement abordé dans la section "I.3. Transfert de fonction et stratégie d'humanisation").

II.5. Conclusions et observations sur la méthodologie de transfert de feuillets

II.5.1. Limitations mises en avant expérimentalement

Le bilan des données expérimentales présentées dans ce chapitre témoigne des difficultés rencontrées lors du transfert de feuillets anti-GFP de Nanofitines, découvertes *in vitro*, à des charpentes protéiques humaines. Dans le cas des structures 1QNT et 1LPJ, l'introduction des 11 acides aminés mutés à la surface des Nanofitines anti-GFP C6 et C8 a pour effet d'induire une insolubilité irréversible dans les conditions que nous avons explorées. Ces mêmes mutations sur la structure 4F7H n'empêchent pas son expression et sa purification sous forme soluble, mais ne lui procurent pas une capacité détectable de fixation spécifique à la GFP. Néanmoins, le transfert de ces résidus fonctionnels de la charpente de Sac7d à celle de Sso7d autorise une fixation spécifique des variants générés, avec une affinité peu altérée. Ce constat pourrait donc suggérer que les propriétés des protéines extrémophiles (comme Sac7d ou Sso7d) leur confèrent une meilleure tolérance à l'introduction d'un nombre important de mutations simultanées. De ce fait, l'exploration de ce groupe de charpentes protéiques représente une piste d'intérêt pour la greffe inter-protéines de résidus clés, mais sort du cadre initial d'une stratégie d'humanisation.

Au delà de ce transfert réussi vers Sso7d, cette étude a mis en avant des obstacles à la capacité de fixation ou à la solubilisation de charpentes humaines devant exposer les feuillets anti-GFP issus de deux Nanofitines différentes, ce qui pourrait suggérer un dénominateur commun à ces obstacles puisque les séquences et les épitopes ciblés sont distincts entre les feuillets transférés. Une analyse plus approfondie a donc été réalisée *a posteriori* pour identifier de possibles biais expliquant ces résultats, et éduquer notre regard d'expérimentateur à la finesse nécessaire pour ce type d'ingénierie protéique.

II.5.2. Prédictions complémentaires implémentées *a posteriori*

Lors de la mise en place de la stratégie de greffe de site de fixation, nous avons criblé les charpentes protéiques hôtes et sélectionné trois d'entre elles, notamment pour leur capacité à

pouvoir être exprimées sous forme recombinante en amont de la résolution de leur structure. En parallèle, nous avons sélectionné deux Nanofitines anti-GFP comme candidats au transfert, en nous focalisant principalement sur les critères de l'épitope ciblé et surtout de l'affinité, anticipant une perte significative de la capacité de fixation après transfert. Au regard des données expérimentales, l'impact du transfert sur les paramètres d'expression des charpentes hôtes a quant à lui été sous estimé lors du criblage initial.

Une étude bioinformatique a donc été menée *a posteriori* afin d'identifier si certains déterminants pouvaient prédire les problèmes d'expression observés expérimentalement. Ces prédictions ont essentiellement porté sur la détermination des différences d'hydrophobie, de stabilité et d'agencement stérique avant ou après greffe des feuillets anti-GFP issus des Nanofitines découvertes par sélection *in vitro*.

II.5.2.1. Hydrophobie

Les interactions hydrophobes peuvent fortement contribuer à la formation d'interfaces protéine-protéine (Chothia et Janin, 1975), ce qui peut nécessiter la présence de régions hydrophobes à la surface exposée au solvant des protéines engagées dans des complexes. Cette participation hydrophobe n'est cependant pas obligatoire dans la stabilisation de complexes, alors qu'elle semble dominante au niveau des cœurs apolaires enfouis, indispensables au bon repliement des protéines (McCoy et al., 1997; Tsai et al., 1997; Xu et al., 1997; Chen et al., 2013). Par conséquent, le transfert d'un site de fixation peut générer un changement de la balance d'hydrophobie au sein de la protéine hôte ou la création d'une région localement plus hydrophobe interférant avec son repliement. Une sur-exposition en surface de résidus aux groupements fonctionnels hydrophobes est d'ailleurs une des causes majeures de l'insolubilité des protéines. Ce paramètre apparaît donc critique pour assurer le maintien de la structure des charpentes hôtes, mais ne conditionne pas nécessairement l'établissement d'une interaction affine. Partant de cette observation et des résultats expérimentaux obtenus, l'analyse des perturbations d'hydrophobie induites par le transfert de feuillet anti-GFP semble pertinente et nous avons tiré avantage des échelles d'hydrophobie publiées par Trinquier et Sanejouand (Trinquier et Sanejouand, 1998) pour évaluer

les différences d'hydrophobies à la surface des Nanofitines anti-GFP identifiées par sélection *in vitro*. La comparaison des scores obtenus (**Figure 46**), proportionnels au potentiel hydrophobe des protéines envisagées, semble indiquer que le feuillet anti-GFP de C8 est un des moins hydrophobes parmi les 8 Nanofitines caractérisées en détails. C'est également, de manière logique, un des feuillets induisant une moindre augmentation d'hydrophobie sur les charpentes insolubles 1QNT et 1LPJ.

		Feuillet anti-GFP greffé							
		E2	D8	C8	B4	C5	C6	C2	A3
Charpente hôte	1AZP	-20	-20	-7	-5	-4	-1	+14	+15
	1LPJ	+3	+3	+16	+18	+19	+22	+37	+38
	1QNT	+7	+7	+20	+22	+23	+26	+41	+42

Figure 46: Scores de variation d'hydrophobie prédite lors des transferts. La différence de score après introduction des résidus fonctionnels des feuillets anti-GFP et les charpentes sauvages a été déterminée, définissant les feuillets induisant une augmentation d'hydrophobie (score > 0) ou une augmentation d'hydrophilie (score < 0) par rapport à la protéine sauvage. Les résidus mutés ont été classés des plus hydrophobes (score maximum) vers les plus hydrophiles (score minimum) dans l'ordre WCMFILVGRS ATP EDKNQHY, d'après une échelle d'hydrophobie initialement proposée lors d'une étude de l'impact de mutation ponctuelles de bases nucléotidiques sur la conservation de cette propriété (Trinquier et Sanejouand, 1998).

Le feuillet identifié sur la Nanofitine anti-GFP C6, quant à lui, présente un profil d'hydrophobie plus important, ce qui aurait pour conséquence une augmentation significative de l'hydrophobie des structures 1QNT ou 1LPJ. Partant de cette hypothèse, nous pouvons nous demander si un feuillet anti-GFP autre que celui de C6 aurait pu être plus approprié au transfert, comme celui de D8 ou E2 qui semblent partager le même épitope à la surface de la GFP.

Ces données confirment que des profils d'hydrophobies variés peuvent être observés pour différentes protéines d'affinité, y compris dans le cas de Nanofitines ciblant des épitopes communs ou chevauchant d'une même protéine. Si l'hypothèse selon laquelle la réduction d'hydrophobie en surface peut effectivement réduire l'insolubilité liée aux mutations des structures 1QNT et 1LPJ était avérée, alors le feuillet anti-GFP de C8 aurait tout de même été sélectionné comme candidat au transfert de charpente. Néanmoins, une Nanofitine autre que C6 aurait pu être considérée pour identifier un feuillet anti-GFP se liant à un épitope distinct de celui fixé par la Nanofitine C8.

II.5.2.2. Stabilité

Il est souvent admis que les protéines observées dans la nature sont le produit d'une optimisation liée à l'évolution, aboutissant à une stabilité suffisante pour assurer leur fonction (Bloom *et al.*, 2006; Tokuriki *et al.*, 2008). Introduire 11 mutations simultanément, en dehors d'un processus d'évolution dirigée, présente donc un risque d'altérer la stabilité des mutants. Une évaluation de l'impact du transfert des feuillets anti-GFP de Nanofitines sur les structures 1AZP, 1C8C, 4F7H, 1QNT et 1LPJ a donc été réalisée à l'aide d'une méthode de design protéique sur squelette peptidique fixe, développée dans le logiciel Rosetta (Dantas *et al.*, 2003). Au regard des scores d'énergie calculés pour les différentes charpentes (**Figure 47**), non mutées ou greffées du feuillet anti-GFP de chacune des 8 Nanofitines caractérisées (A3 à E2), il apparaît que les acides aminés les moins déstabilisants (générant les scores les plus négatifs) sont effectivement ceux des protéines sauvages. A l'inverse, une déstabilisation de plus ou moins grande amplitude a été observée lors de l'introduction des mutations correspondant aux feuillets anti-GFP.

		Feuillet anti-GFP greffé								
		Aucun	C8 ^G	B4	C2	C6 ^G	D8	A3	C5 ^P	E2 ^P
Charpente hôte	1AZP	-126,70	-118,09	-121,74	-113,49	-119,29	-70,43	-118,20	-106,20	34,99
	1C8C	-88,24	-81,08	-83,99	-78,80	-83,45	-41,20	-77,37	-69,03	44,33
	1LPJ	-215,07	-213,23	-201,64	-197,84	-191,60	-202,41	-209,56	121,97	78,12
	1QNT	-230,30	-219,43	-183,48	-211,64	-182,98	-128,83	-99,56	-65,26	-39,94
	4F7H	-152,10	1884,09	-134,96	1871,08	-149,02	-116,02	-143,21	230,10	76,28

Figure 47: Scores de stabilité prédite lors des transferts, calculés avec Rosetta. Les protéines calculées comme étant les plus stables sont figurées en bleu, et les plus instables en rouge. Les scores extrêmement positifs, attribués à un conflit stérique, sont figurés en noir. Les feuillets indiqués par ^P ou ^G représentent respectivement les feuillets anti-GFP contenant une proline ou ayant été greffés expérimentalement.

De manière générale, les feuillets les plus déstabilisants sont ceux de E2 et C5, d'après les calculs effectués sur Rosetta. Cette observation est applicable à l'ensemble des charpentes considérées, y compris la charpente d'origine des Nanofitines. L'analyse des scores individuels des acides aminés attribue ces augmentations de score d'énergie à l'introduction de résidus proline sur E2 et C5, respectivement en sixième et dernière mutation du feuillet (**Figure 31**). La proline présente en

effet des particularités intégrées aux fonctions de scores de Rosetta, puisque l'azote du groupement amine participe à la cyclisation de la chaîne latérale, ne formant pas de liaison hydrogène et présentant ainsi un motif de liaison covalente différent des autres acides aminés, ce qui peut induire des modifications de conformation du squelette peptidique. De ce fait, ces deux candidats présenteraient un risque de modification de l'agencement spatial du squelette de leur feuillet anti-GFP, s'éloignant potentiellement de celui criblé à partir de la structure de Sac7d native, et ne seraient donc pas à privilégier dans la stratégie de transfert de domaine rigide présentée dans ces travaux.

Nous avons également constaté des scores positifs, signifiant une extrême déstabilisation, dans le cas du transfert *in silico* du feuillet de C2 ou C8 sur la structure 4F7H. Dans les deux cas, cette augmentation drastique de score est imputée au premier tryptophane du feuillet anti-GFP, commun à C2 et C8. En effet, il s'avère que les structures prédites de ces mutants de 4F7H présentent un conflit stérique entre le tryptophane introduit (W426) et la chaîne latérale de l'isoleucine I445, ce qui serait une piste explicative de l'absence de fixation observée lors du transfert du feuillet anti-GFP de C8 sur 4F7H. Par ailleurs, l'expression sous forme soluble de ce mutant de 4F7H a été possible, ce qui peut laisser penser que des modifications importantes de structures (locales ou globales) ont éventuellement pu être générées lors de ce transfert pour compenser l'éventuel conflit stérique avec le tryptophane greffé. Cette réorganisation potentielle est une des pistes pouvant expliquer l'absence d'activité de liaison à la GFP des variants de 4F7H.

Ces éléments ne proposent pas d'explication directe à l'insolubilité après transfert du feuillet anti-GFP de C6 ou C8 sur les charpentes ILPJ et IQNT, mais cette analyse a tout de même permis de mettre en évidence un conflit stérique qui n'avait pas été détecté initialement.

II.5.2.3. Diversité des profils prédits

Une variabilité est donc observée parmi les profils d'hydrophobie et de stabilité prédits pour les différentes Nanofitines caractérisées. A l'exception du possible conflit stérique mis en évidence lors du transfert sur 4F7H, l'introduction du feuillet de C8 semblerait induire de faibles déstabilisations des autres charpentes protéiques selon les prédictions de Rosetta, présentant

même les scores les plus favorables pour la mutation des structures 1QNT et 1LPJ. Quant à la Nanofitine C6, son feuillet semble également peu déstabilisant pour l'ensemble des structures considérées, aboutissant au score le plus favorable des mutants de 4F7H.

L'ensemble de ces résultats soulève des questions relatives au nombre considérable de mutations directes à introduire dans une charpente protéique. Était-il possible d'identifier un feuillet anti-GFP qui puisse être adapté sur chaque charpente sélectionnée? Adapter à la fois la charpente et le domaine à y transférer augmenterait-il les chances de succès? Si oui, faut-il générer des sites de fixation plus nombreux et divers, ou cribler un nombre plus important de charpentes? Les feuillets des Nanofitines A3 et D8 pourraient illustrer cette possible nécessité d'adaptation à effectuer par couple domaine de fixation / charpente hôte, puisque leurs scores de stabilité les classent comme peu perturbants (notamment sur 1LPJ) à largement déstabilisants (en particulier sur 1QNT) selon leur charpente hôte.

L'ensemble des prédictions réalisées ont également été effectuées sur les structures 1QNT et 1LPJ à partir de l'ensemble des feuillets anti-GFP séquencés au cours de la sélection de Nanofitines par *ribosome display* (**Figure 48**). Dans les deux cas, C8 reste parmi les candidats les plus intéressants bien que d'autres clones qui n'avaient pas été caractérisés finement auraient pu être utilisés.

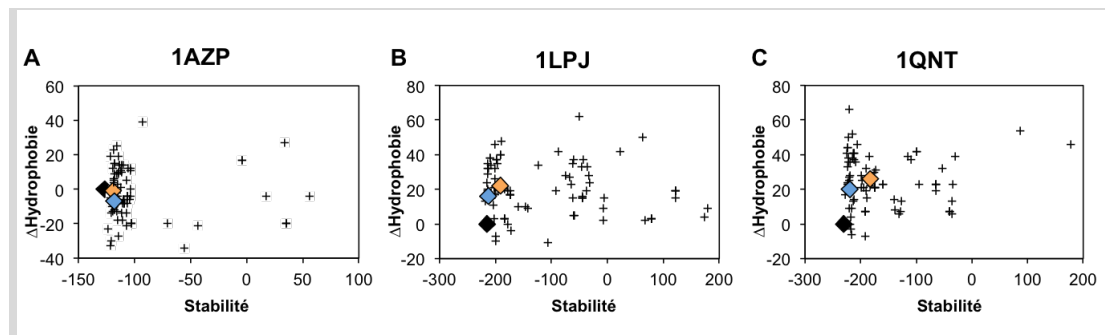


Figure 48: Comparaison de l'ensemble des séquences de feuillets anti-GFP découverts pendant la sélection *in vitro*, en fonction de leur variation d'hydrophobie et de leur stabilité prédites. Les protéines non mutées, greffées du feuillet de C8 ou greffées du feuillet de C6 sont figurées, respectivement, en losanges noirs, bleus et oranges.

II.5.2.4. Cartographie de l'interaction Nanofitine C8 – GFP

D'une part, les études prédictives d'hydrophobie et de stabilité sembleraient indiquer des résultats parmi les plus favorables au transfert du feuillet anti-GFP de C8 sur les structures 1QNT et 1LPJ, qui se sont pourtant soldées expérimentalement par des constructions insolubles, révélant un

problème que nous avons imputé à une hydrophobie ou instabilité trop importantes. Il pourrait être intéressant de réduire la contrainte imposée par l'introduction simultanée de 11 mutations dans de tels cas d'insolubilité, notamment en diminuant le nombre d'acides aminés à greffer.

D'autre part, la bonne expression des mutants de 1C8C et 4F7H démontre un impact moindre sur leur repliement et leur solubilité. Le transfert effectué sur la protéine homologue de Sac7d a permis de conserver une activité de fixation spécifique et affine, tandis que celui sur 4F7H n'a pas démontré de liaison à la GFP, hypothétiquement du fait d'une gêne stérique au niveau du premier résidu du feuillet anti-GFP. Il serait donc intéressant de connaître plus précisément la contribution respective de chaque acide aminé du feuillet anti-GFP de C8 pour approfondir l'étude de ce cas d'humanisation de Nanofitine.

De ce fait, il nous a semblé pertinent d'identifier les résidus cruciaux à l'interaction via la Nanofitine C8 et ainsi cartographier sa surface d'interaction avec la GFP, avec pour objectif de réduire le nombre d'acides aminés à greffer et d'évaluer l'impact du conflit stérique suspecté lors du transfert sur 4F7H. Cette caractérisation a été réalisée par une approche de cartographie par mutations en alanine (ou *alanine scanning*; *Bostrom et al., 2009*), une technique couramment utilisée en biologie moléculaire pour déterminer la contribution d'un résidu spécifique à la stabilité ou à l'activité d'une protéine. Les résidus ainsi étudiés sont mutés en alanine du fait de son groupement fonctionnel méthyl, peu encombrant stériquement et chimiquement inerte, qui peut néanmoins mimer les préférences de structure secondaire que la plupart des autres acides aminés possèdent. Nous avons ainsi réalisé cette étude par l'introduction de mutations ponctuelles en Alanine sur les résidus en surface du feuillet anti-GFP de C8, suivie de mesures de leur impact en ELISA avec des gammes de concentration des mutants Alanine (**Figure 49**).

Les résultats observés, reportés sur la structure tridimensionnelle de Sac7d (**Figure 50A**), ont montré que les mutations des W21A et L26A, respectivement au début et à la fin du premier brin beta du feuillet anti-GFP, impactent drastiquement la capacité de liaison à la GFP. En effet, la concentration nécessaire à l'obtention d'un signal ELISA équivalent à 50% d'un signal saturant en présence de GFP (EC_{50}) est passée de 0,55 nM, pour la Nanofitine C8 non mutée, à plus de 30 et

70 μM , respectivement pour ses mutants W21A et L26A. Un effet moindre a également été observé avec les mutations T31A et L40A, présentes sur les brins 2 et 3 du feuillet anti-GFP, décalant la valeur observée d' EC_{50} au delà de 50 nM. Enfin, la mutation E31A a induit une légère perte d'affinité, traduite par une valeur d' EC_{50} de 2 nM. Les autres positions étudiées n'ont pas modifié significativement les valeurs d' EC_{50} , mais leur implication dans l'interaction avec la GFP ne peut pas être exclue du fait de leurs propriétés. En effet, deux résidus alanine et deux résidus glycine ont été sélectionnés à la surface de la Nanofitine C8, ce qui compromet l'utilisation de la méthode de cartographie par mutations alanine pour ces positions.

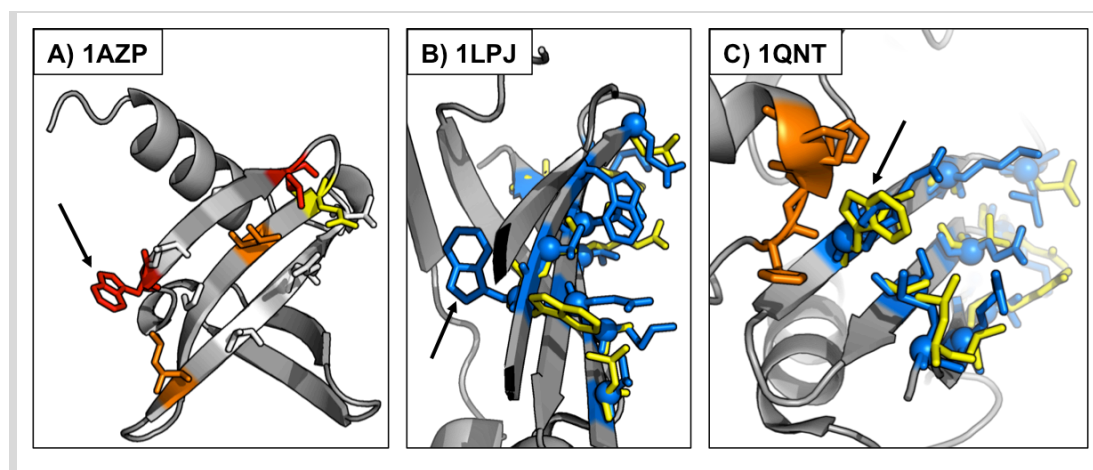
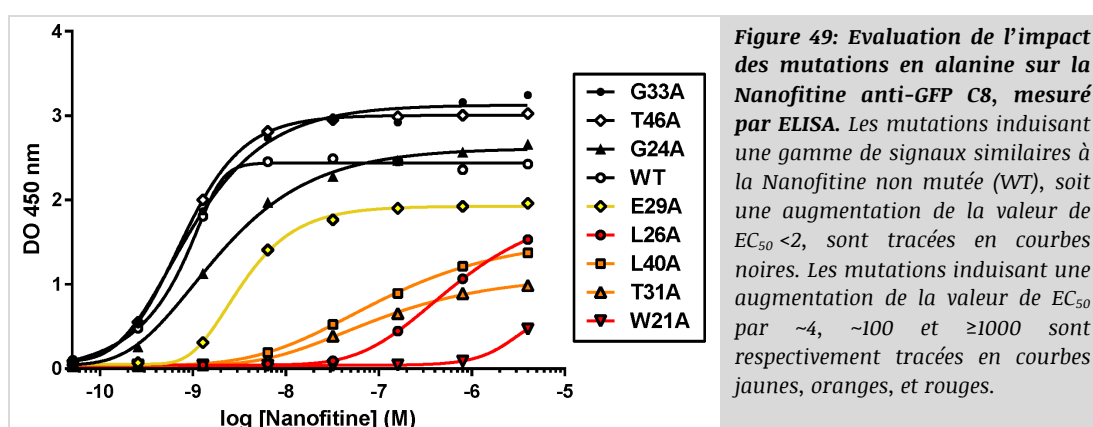


Figure 50: Représentations en cartoon des structures de 1AZP (A), 1LPJ (B) et 1QNT (C). Les résidus représentés en bâtonnets correspondent aux sites de mutations par rapport aux protéines natives. Le résidu désigné par une flèche représente l'emplacement attendu pour le premier tryptophane du feuillet anti-GFP de C8. A) Représentation schématique de l'impact des mutations ponctuelles en alanine sur la Nanofitine C8. Les résidus augmentant drastiquement (~ 1000), fortement (~ 100), moyennement (~ 4) ou de manière négligeable (< 2) la concentration de Nanofitine pour obtenir 50% de signal en ELISA sont représentés, respectivement, en rouge, orange, jaune et blanc. B) et C) Représentation schématique du feuillet anti-GFP de C8 (en jaune) superposé au feuillet à muter (en bleu). Les résidus de surface avec des chaînes latérales à moins de 6 Å du premier tryptophane du feuillet de C8 sont indiqués en orange.

En reportant ces résultats sur la structure de Sac7d (**Figure 50A**), il apparaît donc que chaque brin du feuillet semble impliqué significativement dans l'interaction avec la GFP, aussi bien aux extrémités qu'au centre de la surface rendue aléatoire lors de la sélection par *ribosome display*. Dans le cas de la Nanofitine C8, il aurait donc été difficile de réduire fortement le nombre de positions du feuillet anti-GFP à transférer sur une autre charpente protéique. Nous avons également constaté l'implication cruciale du résidu W21 dans l'interaction C8–GFP, situé à l'extrémité du premier brin du feuillet anti-GFP. Or, l'analyse détaillée des structures 4F7H, 1QNT et 1LPJ souligne des particularités autour de la position où ce tryptophane a été transféré.

Tout d'abord, il apparaît que la mutation C426W, correspondant à l'introduction du tryptophane de C8, génère un conflit stérique avec le résidu I445, tel que discuté dans la section "II.5.2.2. Stabilité". Bien que n'empêchant pas l'expression du mutant 4F7H avec le feuillet anti-GFP de C8, une gêne stérique pourrait être à l'origine de l'absence de fixation à la GFP. Par ailleurs, l'introduction de ce résidu sur la structure 1LPJ correspond à une mutation silencieuse, ce qui pourrait suggérer un moindre effet sur la stabilité du mutant. Pourtant, l'analyse de la structure tridimensionnelle 1LPJ révèle que ce résidu tryptophane n'est pas exposé à la surface de la protéine mais enfoui (**Figure 50B**). Par conséquent, les données générées par la cartographie de C8 nous permettent de conclure que le transfert du feuillet anti-GFP de C8 sur 1LPJ aurait fortement impacté sa capacité à fixer la GFP, s'il avait été produit sous forme soluble. Enfin, bien que devant être exposé à la surface sans générer de conflit stérique, ce tryptophane ne semble pas non plus disposer d'un environnement favorable après transfert sur la structure 1QNT (**Figure 50C**). Si ce mutant avait été exprimé sous forme soluble, le tryptophane de C8 y serait potentiellement partiellement masqué par le motif HHP des positions 85 à 87 de la protéine hôte, dont les chaînes latérales se situent au voisinage direct du tryptophane après transfert (à moins de 6 Å).

L'ensemble de ces observations, complétant les données expérimentales accumulées sur l'insolubilité ou l'absence de fixation à la GFP des protéines humaines greffées du feuillet anti-GFP de C8, vont dans le sens d'une incompatibilité entre ce feuillet anti-GFP et les charpentes

hôtes identifiées. Pour augmenter les chances de succès de tels transferts sur des charpentes plus complexes que celles de Sac7d et Sso7d, il serait donc pertinent d'anticiper l'acquisition de ces informations tout en profitant d'optimisations sur le plan expérimental.

II.5.3. Conclusions et perspectives d'améliorations méthodologiques

Lors de la mise en place de cette stratégie originale d'humanisation de protéines de fixation, basée sur la découverte de domaines de fixations sur la charpente des Nanofitines à greffer sur des protéines d'origine humaine, nous nous sommes heurtés à l'impossibilité de transférer efficacement la capacité de liaison d'affinité nanomolaire à la GFP. Bien que décevant, ce constat souligne l'importance et la plus-value des outils de modélisation pour l'ingénierie protéique, accompagnés d'un regard éduqué de l'expérimentateur.

L'expérience acquise lors des manipulations expérimentales et des analyses bioinformatiques réalisées a mis en évidence des pistes pour augmenter les chances de succès d'une telle approche. Tout d'abord, nous pouvons noter que le premier obstacle qui n'a pas été contourné expérimentalement concerne l'insolubilité induite lors du transfert de feuillets anti-GFP complets. Qu'il s'agisse de problèmes de repliement ou de précipitation de protéines dans une conformation correcte, les travaux présentés dans ce chapitre suggèrent que l'utilisation des données structurales pourrait anticiper ces phénomènes. Notamment, ces analyses devraient aboutir à la prédiction d'éventuels conflits stériques ou mauvaise orientation d'acides aminés à muter. Une des améliorations envisageables consisterait ainsi à affiner l'évaluation de l'environnement des résidus à transférer, réalisé ici par calcul des similarités cosinus des vecteurs N→C et estimation des variations de RMS d'accessibilité, tout en imposant un contrôle par l'expérimentateur lors des dernières phases de criblage. Dans cette étude, ceci aurait permis de déceler précocement des incompatibilités de domaines anti-GFP vis-à-vis de l'environnement dans lequel le transfert a été effectué.

Par ailleurs, des facteurs pourraient rendre le défi d'un transfert de feuillet complet plus accessible. Notamment, les progrès techniques constants qui facilitent les manipulations de gènes (*van den Ent et Löwe, 2006; Gibson et al., 2010*) permettraient de réaliser des études

systématiques de cartographie des Nanofitines par mutations alanine, rendant possible la découverte de domaines de fixation de dimension plus restreinte imposant donc moins de contraintes. De même, les mutations pourraient être transférées progressivement à la surface des charpentes à greffer, et non pas d'un unique bloc. La complémentarité de ces pistes d'amélioration pourrait ainsi donner accès à des protéines avec des profils de stabilité et de solubilité plus proches de ceux des protéines sauvages que des mutants décrits dans ce chapitre.

Un élément primordial de la méthodologie employée ici est que la découverte des charpentes hôtes est basée sur les structures résolues disponibles dans la PDB. Pourtant, il est très probable que des variations de repliement, d'importance variable, puissent apparaître lors de l'introduction de mutations, même si celles-ci sont opérées en surface. Ceci s'applique aux charpentes humaines testées, mais également à la charpente des Nanofitines dérivées de Sac7d comme en témoigne la variabilité observée pour les structures des mutants de Sac7d cristallisés (principalement dans les boucles, mais également aux extrémités de l'hélice ou des feuilletts, **Figure 51**). L'accès aux structures résolues des Nanofitines dont le domaine de fixation est à transférer, idéalement co-cristallisées avec leur cible, pourrait limiter ce phénomène et permettre d'ajuster les paramètres de criblage lors de la recherche de charpente hôte. Ceci permettrait également de vérifier si des acides aminés non mutés dans les Nanofitines sont impliqués dans l'interface formée. Toutefois, de telles structures ne proposeraient que peu d'indications concernant les modifications induites sur la charpente hôte.

Se pose également la question de la capacité des charpentes hôtes à subir un nombre considérable de mutations simultanées. D'une part, cette problématique nécessite sans doute d'être abordée avec plus de finesse, en adaptant par exemple le criblage des charpentes pour chaque domaine de Nanofitine à transférer. Il est notamment envisageable de ne plus cribler les structures protéiques seulement sur des paramètres géométriques, mais en tenant également compte des propriétés des groupements fonctionnels de la surface d'interaction à greffer pour limiter la déstabilisation de la charpente hôte.

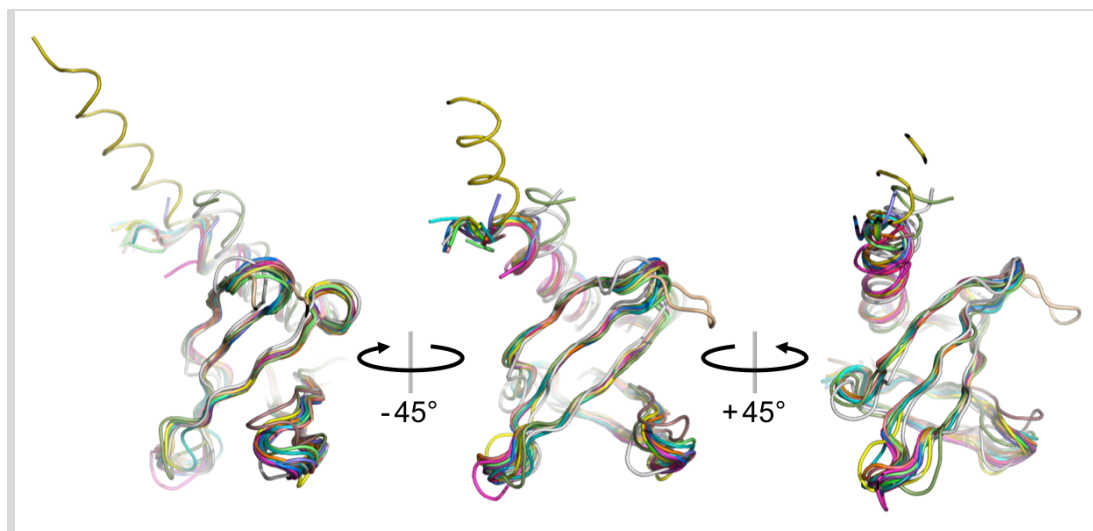
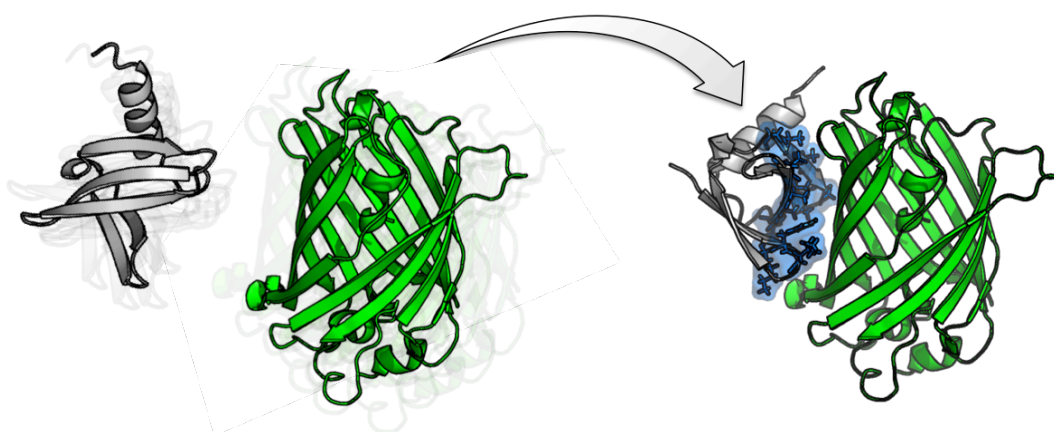


Figure 51: Superposition des structures de variants de Sac7d disponibles dans la PDB. Le squelette peptidique de chaque chaîne protéique est représenté d'une couleur différente. Identifiants PDB: 1AZP, 1AZQ, 1BF4, 1CA5, 1CA6, 1SAP, 1WDO, 1WDI, 1WTO, 1WTP, 1WTQ, 1WTR, 1WTV, 1WTW, 1WTX, 1WVL, 1XX8, 1XYI, 2XIW, 4CJI, 4CJZ.

D'autre part, nous avons postulé que l'humanisation des Nanofitines pouvait répondre à une nécessité pour leur utilisation thérapeutique, lors du démarrage des travaux présentés dans ce manuscrit. Cependant, les données générées au cours des trois dernières années à ce propos tendent à infirmer ce besoin, suggérant des profils pharmacocinétiques ou immunogéniques favorables malgré l'origine archéobactérienne des Nanofitines. Il serait donc envisageable d'ouvrir le choix des protéines hôtes criblées et ainsi choisir des charpentes plus résistantes sans leur imposer une origine humaine, à condition que leur profil réponde aux impératifs d'une utilisation thérapeutique.

Pour finir, nous pouvons noter que cette étude complète l'étalonnage de la difficulté à réaliser rationnellement le transfert d'acides aminés d'une protéine à l'autre. En dehors des problématiques soulevées lors du transfert sur les charpentes humaines 4F7H, 1QNT et 1LPJ, nous avons pu mettre en évidence l'impact du passage de la charpente de Sac7d à son celle de son homologue Sso7d sur l'activité de Nanofitines. Malgré des séquences et des structures tri-dimensionnelles très proches, nous avons pu observer une légère diminution d'affinité, principalement liée à une augmentation de la cinétique de dissociation. Bien que moins importante que celle observée lors d'un transfert de feuillet anti-IgG avec boucles (Béhar *et al.*, 2014), cette perte d'affinité souligne la précision nécessaire pour entreprendre un tel transfert.

Chapitre III: Design de novo de Nanofitine spécifique



Chapitre III: Design *de novo* de Nanofitine spécifique

III.1. Stratégie de design *de novo* de Nanofitines anti-GFP

III.1.1. Complémentarité des approches *in silico* et *in vitro*

Le développement de protéines d'affinités à usage biotechnologique ou médical nécessite le contrôle efficace et précis de l'interaction entre ces molécules et leur cible. D'une part, la rationalisation de ce développement a fait l'objet de travaux récents (décrits en introduction dans la section "I.4. Approches computationnelles de design rationnel"), démontrant des succès ponctuels du design computationnel d'interfaces protéiques. Encore jeunes, ces approches ne sont pas limitées par la diversité de séquences différentes à évaluer et représentent des avancées encourageantes des méthodes prédictives. Toutefois, ces prouesses techniques sont mitigées par leur dépendance persistante aux techniques de maturation d'affinité *in vitro* malgré l'utilisation courante de plusieurs charpentes protéiques à muter, et par la proportion modérée des complexes prédits confirmés expérimentalement. Ce constat réside dans le fait que les méthodes développées *in silico* souffrent de fonctions de scores imparfaites à l'heure actuelle, ainsi que d'un espace conformationnel à explorer extrêmement large. D'autre part, la sélection *in vitro* de Nanofitines ne présente pas ces biais de modélisation et représente une solution pragmatique aboutissant généralement à des molécules d'affinité nanomolaire. Cependant, leur sélection par *ribosome display* flirte avec les limites de diversité théoriques de la technique de sélection *in vitro*, tout en étant plus difficilement dirigée avec précision vers des épitopes prédéfinis. Ces contraintes nécessitent donc une capacité de criblage avec un débit important en aval du processus de sélection, représentant un effort croissant dans le cas des stratégies de ciblage les plus élaborées (présentant le plus souvent un risque d'enrichissement sur des régions non désirées). La combinaison de ces deux types d'approches, qui apparaissent complémentaires, présentent donc un attrait manifeste pour guider la découverte de Nanofitines spécifiques d'épitopes choisis par avance (de la génération de banques conçues rationnellement à la prédiction de variants ponctuels affins).

Ainsi, nous avons conçu une approche de design *de novo* de Nanofitines s'appuyant sur les compétences du logiciel Rosetta afin d'améliorer la rationalisation de la génération des Nanofitines. Bien que nécessitant une expertise des utilisateurs et une adaptation à chaque situation (comme souligné dans les comparaisons des groupes humains et des serveurs de prédiction dans les projets CASP et CAPRI), nous avons constaté que de nombreux outils ont été intégrés à Rosetta. De plus, la majorité de ces applications a été documentée à travers les publications de journaux scientifiques ainsi que via une documentation maintenue par les auteurs et la communauté d'utilisateurs de Rosetta pour être mise à disposition de la communauté scientifique (Kaufmann *et al.*, 2010).

III.1.2. Buts et stratégie de design de Nanofitines

Le but de notre étude de design consiste à modéliser la formation d'un complexe stable entre une Nanofitine et sa cible via une interface déterminée précisément, réduisant ainsi l'effort de criblage nécessaire à l'identification d'une Nanofitine visant un épitope donné. Pour cela, nous avons évalué notre capacité à prédire des séquences de variants (ou des ensembles de séquences pouvant guider la conception de banques dirigées) dans le cadre d'une preuve de concept centrée sur le design *de novo* de Nanofitines dirigées contre la GFP. Les critères de réussite de notre approche résident finalement dans deux prédictions: déterminer si des Nanofitines peuvent cibler un épitope défini à l'avance (critère majoritairement géométrique), et identifier des profils de séquences de Nanofitines affines et spécifiques (design au sens propre). A terme, nous espérons que ce type d'approche pourrait, en cas de succès, modéliser la corrélation entre site de fixation et rôle fonctionnel des Nanofitines, et ainsi nous aider à anticiper et contrôler les actions effectrices des Nanofitines au moment de leur liaison.

Pour répondre à ces objectifs, nous avons utilisé des méthodes d'échantillonnage et des fonctions de calcul d'énergie similaires à celles engagées dans le re-design d'interfaces, avec trois défis méthodologiques supplémentaires. Le premier, majeur, est la génération des orientations de complexes Nanofitine:cible avant design, sans appui sur une interface naturelle déjà existante. Le deuxième, lié à notre volonté d'appliquer cette stratégie aux Nanofitines, est la restriction à la

charpente de Sac7d comme seule protéine d'affinité à muter (contrairement aux exemples actuels les plus prometteurs qui ont exploré des banques de charpentes différentes). De ce fait, cette approche n'a pas bénéficié des éléments évolutifs (tels que des complémentarités de squelette peptidique) qui peuvent prédisposer des partenaires à interagir entre eux, mais pourrait permettre de découvrir ces orientations complémentaires *de novo*. Le dernier réside dans notre objectif de ciblage de protéines dans leur état natif, ce qui implique l'introduction de mutations à la surface des Nanofitines seulement, en interdisant toute mutagenèse de leur partenaire protéique ciblé.

Au cours d'une démonstration centrée sur la GFP et les Nanofitines, nous avons considéré que cette interaction modèle était réalisée par association entre une protéine aux surfaces relativement planes (du fait de sa structure en tonneau beta) et la surface d'interaction naturelle des Nanofitines, essentiellement constituée d'un feuillet beta. De ce fait, les structures des partenaires impliqués dans les complexes à modéliser *de novo* ont été jugées relativement rigides (pour les mêmes raisons que celles invoquées précédemment dans la présentation de notre stratégie d'humanisation, dans la section "I.3.3.4. Transferts de feuillets beta"). Nous avons donc fait le postulat que l'introduction de flexibilité des chaînes principales n'était pas nécessaire lors de cette démonstration axée sur la GFP et les Nanofitines, compte tenu des résultats obtenus via Rosetta avec des structures relativement rigides (principalement illustrées lors des tests CAPRI).

Finalement, la démarche de design *de novo* que nous avons appliquée avec pour objectif le ciblage rationnel de Nanofitines peut être résumée en quatre étapes principales (**Figure 52**). A partir de structures préalablement résolues, des poses initiales ont été générées pour chaque partenaire moléculaire devant former l'interaction (étape 1). Une simulation d'ancrage moléculaire a alors permis d'orienter aléatoirement un des partenaires au contact de la surface de l'autre protéine (étape 2), correspondant respectivement à la Nanofitine et à sa cible dans cette étude. Les meilleurs résultats d'ancrage moléculaire ont ensuite été classés par groupes pour identifier les régions les plus favorables à une interaction de type protéine-protéine à la surface de la cible (étape 3). Finalement, des poses issues des meilleurs groupes ont été utilisées pour introduire des

mutations aléatoires sur la surface de la Nanofitine au niveau de l'interface entre les deux partenaires, afin de générer une Nanofitine spécifique et affine pour la région ciblée (étape 4). Les résultats de ces différentes étapes sont discutés successivement dans ce chapitre, pour ensuite se concentrer sur la construction, l'expression et la caractérisation expérimentale d'une famille de Nanofitines dirigée contre la GFP par design *in silico*.

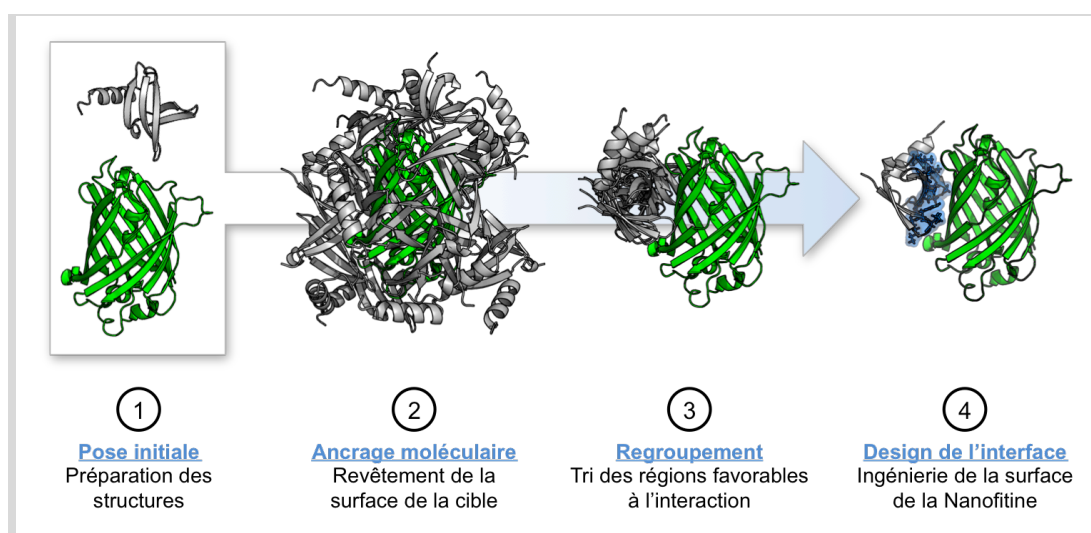


Figure 52: Logigramme des principales étapes du design de novo de Nanofitine. Les structures schématisées en gris et en vert représentent, respectivement, la charpente de Sac7d et la GFP. La représentation en surface et bâtonnets de couleur bleue représente les résidus de la Nanofitine au contact de la GFP.

III.2. Obtention et préparation des poses initiales

Les structures tridimensionnelles utilisées comme poses initiales doivent être choisies avec précaution car elles doivent refléter au mieux l'état d'une Nanofitine et de sa cible lors du processus de prédiction bioinformatique, malgré l'introduction de mutations ou la formation d'un complexe à différentes interfaces. La méthode que nous avons mise en oeuvre trouve son point de départ dans les structures préalablement résolues de Sac7d (identifiant PDB: 1AZP; Robinson *et al.*, 1998) et de la GFP dont nous disposons sous forme recombinante, issue de la méduse *Aequorea coerulescens* (identifiant PDN: 3LVA; Pletneva *et al.*, 2010).

Dans ce type d'analyse prédictive, le choix de la structure qui verra une partie de ses résidus rendus aléatoires lors de l'étape finale de design protéique doit principalement pouvoir tenir compte de l'impact de telles mutations. Nous avons émis l'hypothèse que les Nanofitines peuvent répondre idéalement à cette contrainte, notamment grâce au coeur hydrophobe de Sac7d assurant son extrême stabilité. Cet agencement spatial assurerait un squelette peptidique

relativement fixe, et en particulier dans les régions les mieux organisées telles que le second feuillet beta de Sac7d, comme nous l'avons déjà évoqué dans la section "II.5.3. Conclusions et perspectives d'améliorations méthodologiques" du chapitre dédié à l'humanisation de Nanofitines (**Figure 51**). Ces propriétés de Sac7d seraient également un des facteurs rendant sa mutagenèse aléatoire possible pour la génération de Nanofitines stables, comme largement éprouvé au sein du laboratoire. De ce fait, nous avons supposé qu'introduire des mutations sur Sac7d, principalement dans ce feuillet, affecterait de façon très limitée la conformation des Nanofitines générées par rapport à la structure de Sac7d utilisée lors des prédictions.

D'autre part, la structure de la GFP a été choisie avec soin pour correspondre à la structure la plus proche du variant disponible au laboratoire, qui a été utilisé ensuite pour évaluer la fixation entre les Nanofitines prédites *in silico* et la GFP. De plus, la superposition de la structure 3LVA sur celle d'un complexe Nanobody:GFP (identifiant PDB 3OGO; Kubala *et al.*, 2010) a révélé que la région C-terminale de la protéine, désorganisée, peut masquer une partie de la structure en tonneau beta de la GFP pouvant pourtant être ciblée par une protéine d'affinité (**Figure 53**). Notre choix s'est donc porté sur la structure publiée sous l'identifiant PDB 3LVA, après suppression de son extrémité C-terminale labile. Cette modification devrait ainsi avoir abrogé l'encombrement stérique observé, qui semblerait plutôt résulter d'un artefact de la cristallisation.

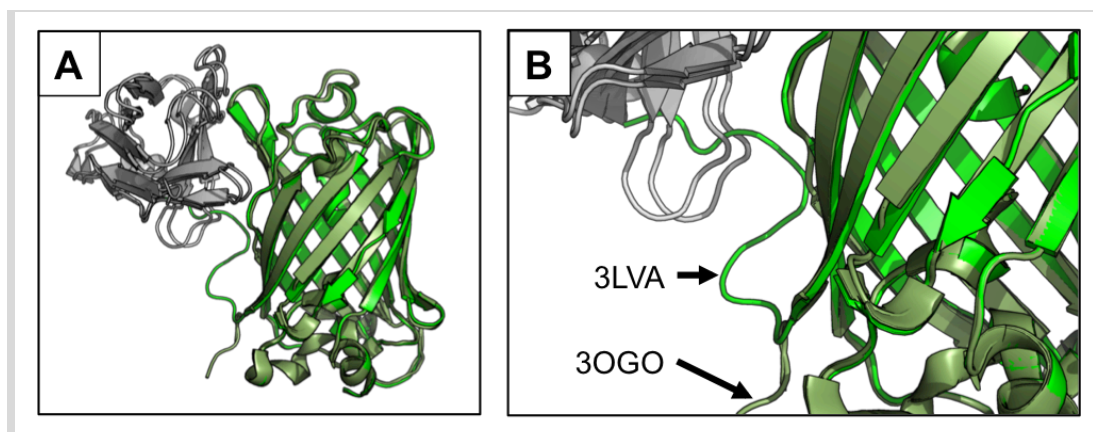


Figure 53: Superposition du complexe Nanobody:GFP 3OGO et de la GFP seule (3LVA). Les structures représentées en gris correspondent au Nanobody en complexe avec la GFP (vert foncé). La GFP de la structure 3LVA, correspondant à la GFP recombinante utilisée *in vitro*, est représentée en gris clair et est superposée à la GFP de la structure 3OGO. Les différences de positionnement de l'extrémité C-terminale labile de la GFP sont indiquées par des flèches.

Après avoir identifié les structures de départ pour cette étude de design protéique *in silico*, une dernière opération a été réalisée avant de les engager dans l'étape d'ancrage moléculaire. Il s'agit d'une minimisation du score calculé par Rosetta effectuée par relaxation tout en conservant les atomes aussi proches que possible de leur position originale dans le cristal, afin de limiter les biais d'attribution de score lors des étapes suivantes. Brièvement, une relaxation a été réalisée en permettant une réorganisation au niveau des chaînes latérales des acides aminés de chaque protéine selon l'ensemble des rotamères possibles, mais également en autorisant des mouvements du squelette peptidique. Contrairement à un protocole de relaxation standard, des contraintes ont cependant été ajoutées pour limiter les mouvements des acides aminés, dont les chaînes latérales sont représentées par des pseudo-atomes (ou centroïdes) au cours de ces simulations. Ce protocole de relaxation contrainte, testé par les auteurs de Rosetta avant design de 51 enzymes, permet de retrouver 5% de séquences en plus par rapport aux structures non modifiées disponibles directement dans la PDB et aboutit en moyenne à un RMSD des carbones alpha de 0,077 Å (données fournies au sein de la suite logicielle).

	1	66
Sac7d	MVKVKFKYKGEEKEVDTSKI	KKVWRVGMVSFTYDDNGKTGRGAVSEKDAPKELLDMLARAEREKK
Ala	MVKVKFKYKGEEKEVDTSKI	AAVARAGKAVAFAYDDNGKAGAGAVA EKDAPKELLDMLARAEREKK
C6	MVKVKFKYKGEEKEVDTSKI	TLVVRFGKLVTFHYDDNGKWTGAVQEKDAPKELLDMLARAEREKK
C6 shuffle	MVKVKFKYKGEEKEVDTSKI	LVARWGKTVFFTYDDNGKHGQGTVLEKDAPKELLDMLARAEREKK
C8	MVKVKYKYKGEEKEVDTSKI	WAVGRLGKEVTFGYDDNGKLGAGRVTEKDAPKELLDMLARAEREKK
C8 shuffle	MVKVKYKYKGEEKEVDTSKI	AGVARTGKLWFTYDDNGKRGLGAVGEKDAPKELLDMLARAEREKK
E2	MVKVKFKYKGEEKEVDTSKI	TSVHRLGKYVPFVYDDNGKTGTGWVDEKDAPKELLDMLARAEREKK
E2 shuffle	MVKVKFKYKGEEKEVDTSKI	PDVLRSGKTVWFTYDDNGKVGHGVT EKDAPKELLDMLARAEREKK

Figure 54: Alignement des séquences des variants de Sac7d utilisées en ancrage moléculaire. Les positions mutées dans le second feuillet beta de la charpente de Sac7d sont encadrées en noir.

Pour finir, des structures alternatives de 1AZP ont été générées grâce au protocole de design avec squelette peptidique fixe de Rosetta (**Figure 54**), par suspicion d'un impact de la séquence des Nanofitines dans les résultats d'ancrage moléculaire à suivre. Ces structures ont été produites en introduisant les mutations des Nanofitines anti-GFP C6, C8 et E2 (dont la découverte a été discutée précédemment dans la section "II.2. Découverte de Nanofitines anti-GFP interagissant via leur feuillet"). Des structures de variants ont également été générées à partir des séquences

mélangées de ces Nanofitines, désignées avec le suffixe “shuffle” dans ce manuscrit. Une dernière structure alternative a été produite en mutant les positions variables entre les Nanofitines par des résidus alanine.

Au total, 8 séquences ont été exposées à la surface de la structure 1AZP relaxée, en comptant Sac7d sauvage et ses 7 variants, pour être fixées par ancrage moléculaire à la structure 3LVA relaxée sans son extrémité C-terminale.

III.3. Ancrage moléculaire pour la recherche de complémentarité spatiale

III.3.1. Description de l'algorithme

L'ancrage moléculaire entre Nanofitines et GFP, respectivement modélisées avec les structures dérivées de 1AZP et de 3LVA, a été effectué à l'aide du protocole RosettaDock (Chaudhury et al., 2011) tel que résumé sur la **Figure 55**. Succinctement, l'algorithme utilisé a été décrit par ses auteurs comme suivant grossièrement la théorie biophysique de la rencontre de deux protéines en complexe suivie d'une transition à un état lié.

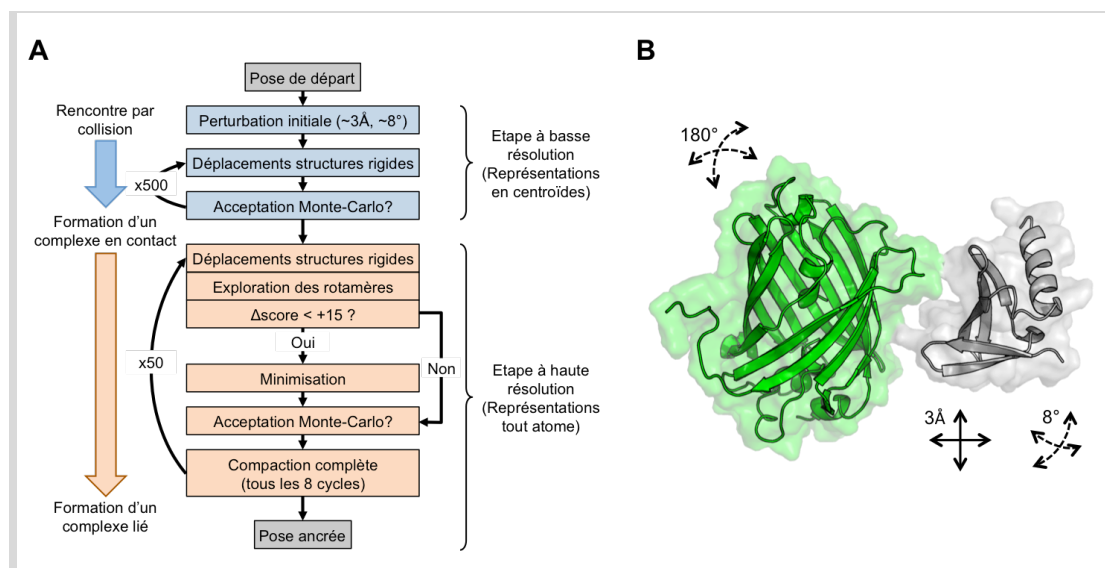


Figure 55: Représentation schématique du protocole d'ancrage moléculaire. A) Logigramme de l'algorithme de RosettaDock, adapté d'après Chaudhury et al., 2011. B) Représentation des perturbations initiales imposées aux structures de la GFP (vert) et de Sac7d (gris) pour la formation d'un complexe.

Pour chaque résultat de simulation obtenu, cet algorithme comporte plusieurs cycles de randomisation basés sur une méthode de Monte-Carlo en conservant un squelette peptidique fixe. Deux types de cycles sont impliqués après une perturbation initiale rendant totalement aléatoire l'orientation de la cible (rotations jusqu'à 180° par axe) et partiellement celle de la

Nanofitine (rotations et translations jusqu'à 8° et 3 Å, respectivement). Dans un premier temps, 500 cycles de perturbations (rotations et translations), dits gros-grain, sont effectués à plus faible résolution jusqu'à formation d'un complexe par la rencontre des deux partenaires moléculaires dont les structures sont simplifiées en remplaçant leurs chaînes latérales par leur centroïde. Dans un second temps, 50 cycles sont effectués à plus haute résolution pour générer les complexes liés à partir du complexe présentant la plus faible énergie, tout en intégrant des étapes additionnelles de minimisation de score. Ces cycles remplacent également les représentations en centroïdes par l'ensemble des atomes des chaînes latérales, ce qui permet d'affiner les simulations en ajoutant l'exploration des rotamères à celle des orientations globales des protéines. En fin de cycles, une pose est finalement retenue à partir du meilleur score d'énergie obtenu, représentant une proposition de complexe possible qui sera désignée dans la suite de ce manuscrit comme pose d'ancrage moléculaire.

Lors de cette étude, nous avons généré 100 000 poses d'ancrage moléculaire entre la GFP et chacune des 8 structures de Sac7d préalablement préparées, soit 800 000 complexes formés. Chaque complexe a ensuite été trié sur la base de son score d'interaction, représentant la différence entre l'énergie du complexe et la somme des énergies des partenaires non complexés. Nous avons ensuite considéré les complexes présentant les scores d'interactions les plus avantageux, typiquement compris entre -5 et -10.

III.3.2. Impact des séquences des Nanofitines

L'algorithme d'ancrage moléculaire de Rosetta ayant démontré des succès pour la reconstitution de complexes protéiques, nous avons souhaité évaluer la robustesse de cet outil en effectuant des simulations à partir de Nanofitines déjà identifiées comme spécifiques de la GFP. Initialement, le choix des 8 séquences utilisées sur la structure 1AZP lors de l'ancrage moléculaire a donc été effectué pour représenter 3 Nanofitines anti-GFP de séquences suffisamment divergentes (comme préalablement décrit dans la section "II.2. Découverte de Nanofitines anti-GFP interagissant via leur feuillet"), aboutissant au choix des clones C6, C8 et E2. Ce choix a également été conforté par la découverte d'épitopes distincts liés par la Nanofitine C8 et par les autres Nanofitines anti-GFP,

tel que discuté plus particulièrement dans la section “II.2.3.4. Identification d'épitopes distincts sur la GFP”. Dans le cas de résultats exacts d'ancrage moléculaire, nous avons émis l'hypothèse que les meilleures poses générées devraient refléter les différences de fixation observées expérimentalement.

Des séquences contrôles ont également été ajoutées à ces prédictions, à savoir la séquence de Sac7d non mutée et un variant de Sac7d exposant des Alanine au lieu des résidus mutés sur les variants Nanofitines anti-GFP. Ces deux conditions ont été prévues pour ne pas pouvoir se fixer à la GFP, et donc aboutir à des résultats d'ancrage moléculaire moins intéressants en matière de scores d'interaction. Des contrôles négatifs plus fins ont également été intégrés via les variants “shuffle” des trois Nanofitines anti-GFP, dont les acides aminés mutés ont été intervertis aléatoirement à la surface de la structure IAZP. Ces dernières conditions ont été ajoutées en prévoyant qu'elles pourraient générer des scores d'interaction moins intéressants, ou sur des régions différentes de la GFP.

Les résultats après simulations d'ancrage moléculaire ont donc été anticipés sur deux aspects. Tout d'abord, nous avons souhaité évaluer la capacité à distinguer efficacement des épitopes différents ciblés par des Nanofitines préalablement identifiées. Plus important encore, nous avons recherché des poses énergétiquement attractives pour fournir des épitopes et orientations favorables à une interaction protéine-protéine en vue du design *in silico* à suivre.

Après avoir généré 100 000 poses d'ancrage moléculaire par condition, les meilleurs résultats ont été filtrés et ont montré une convergence vers des régions spécifiques de la GFP. En considérant tout d'abord les poses ayant les scores d'interaction les plus négatifs avec les Nanofitines anti-GFP (**Figure 56A**), deux groupes ont été identifiés indépendamment de la Nanofitine concernée. Contrairement à nos attentes, les conditions témoins ont également abouti à des résultats d'ancrage moléculaire avec des scores satisfaisants, et leur localisation à la surface de la GFP a semblé proche de celle observée pour les Nanofitines anti-GFP (**Figure 56B**). Cette observation soulève donc des questions concernant l'impact spécifique des résidus en surface de la charpente des Nanofitines dans ces simulations. Ces résultats tendraient à indiquer que la convergence

observée ne provient pas essentiellement de la composition des mutants de Sac7d étudiés, mais plutôt de la nature des groupements fonctionnels à la surface de la GFP ou d'une complémentarité géométrique entre la charpente des Nanofitines et ces épitopes de la GFP.

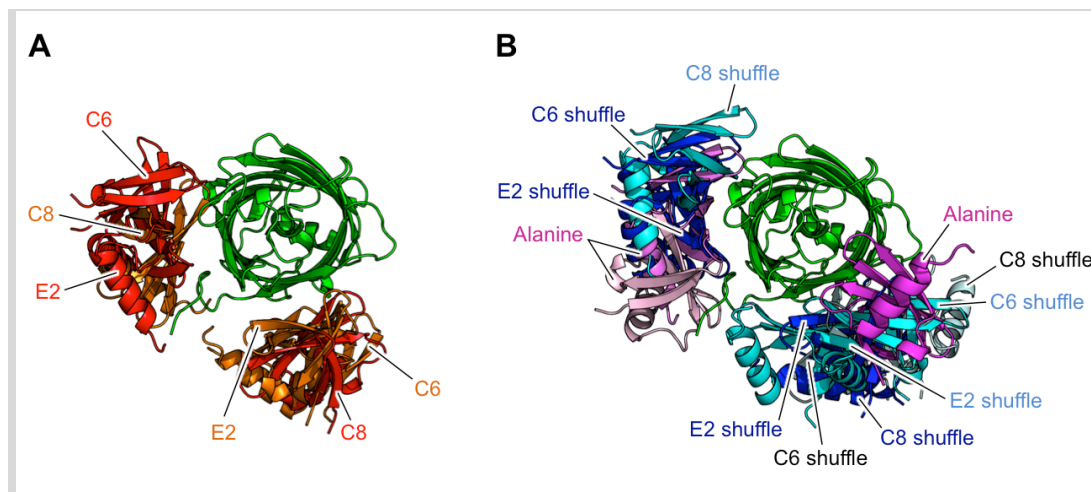


Figure 56: Exemples de complexes obtenus en ancrage moléculaire avec les séquences de Nanofitines anti-GFP (A) ou des Nanofitines non spécifiques (B). Pour chaque séquence utilisée, les structures de Nanofitines au centre des interfaces aboutissant aux meilleurs scores d'interaction sont représentées en cartoon (en couleur), autour de la GFP (en vert).

Bien que les simulations d'ancrage moléculaire aient abouti à une absence de spécificité vis-à-vis de la séquence de la Nanofitine, cet aspect n'a pas été investigué davantage. En effet, ces résultats indiquent un ciblage préférentiel de deux sites d'interaction à la surface de la GFP, ce qui pourrait être en adéquation avec l'objectif principal de cette partie de l'étude: identifier des poses de départ pour le design protéique. L'ensemble des complexes ayant obtenu un score d'interaction inférieur à -5 a donc été caractérisé par la suite afin de définir les poses les plus prometteuses pour la génération d'une interface Nanofitine:GFP.

III.3.3. Découverte de régions et d'orientations géométriques favorables au ciblage

Afin d'identifier les poses d'ancrage moléculaire à utiliser pour le design protéique, nous avons comparé plus finement les poses présentant les meilleurs scores d'interaction en les classant par groupes à la surface de la GFP dans un premier temps. Pour ce faire, le centre de chaque surface d'interaction a été représenté par un pseudo-atome (représenté par une sphère sur la **Figure 57**) puis ces pseudo-atomes ont été répartis en groupes à l'aide du logiciel Gromacs, plus particulièrement via l'application *g_cluster* (Daura et al., 1999). Le seuil de coupure appliqué au regroupement des structures a été fixé à une valeur de 0,03 nm pour normaliser la découverte de

groupes de dimension équivalente et suffisamment restreinte. Un exemple de différents seuils de coupure est représenté sur la **Figure 57A** à partir du groupe privilégié avec la séquence de C6. Pour chaque condition, les groupes comportant le plus de membres avec les meilleurs scores ont ensuite été représentés en centroïdes, comme montré sur la **Figure 57B**, pour être comparés plus facilement. Comme attendu d'après les résultats précédents, nous avons retrouvé deux pôles de la GFP plus particulièrement ciblés par les meilleures poses d'ancrage moléculaire.

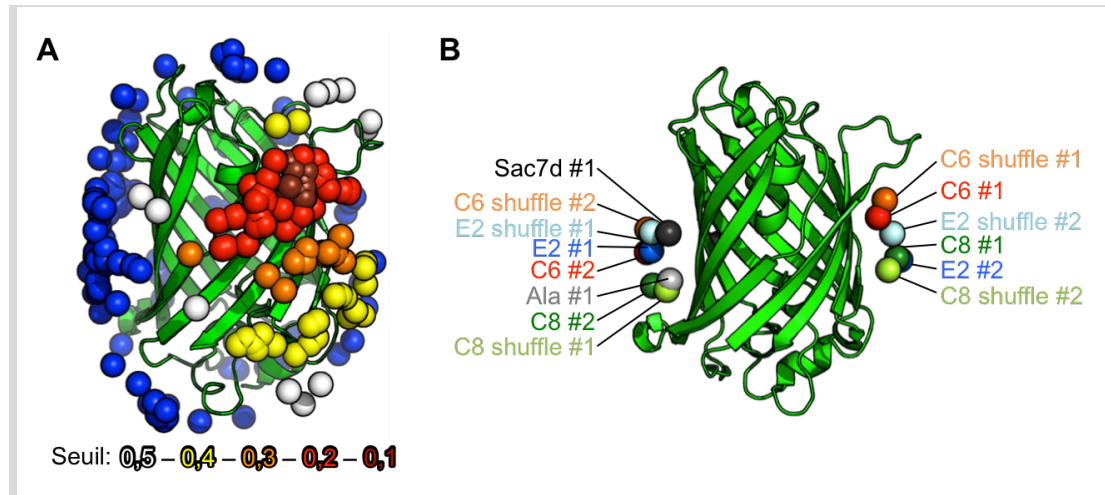


Figure 57: Identification des groupes prédits par ancrage moléculaire à la surface de la GFP (vert). A) Représentation du groupe principal obtenue à partir des poses ayant les meilleurs scores d'interactions avec la séquence de la Nanofitine C6. Le centre de chaque surface d'interaction est représenté par une sphère, colorée en bleu si la pose n'est pas comprise dans le groupe, et colorée en marron, rouge, orange, jaune ou blanc si la pose est comprise dans le groupe (par ordre croissant de seuil de coupure). B) Représentation en sphères des groupes comprenant le plus de poses pour chaque condition explorée, réparties sur deux pôles à la surface de la GFP.

Le premier pôle est composé d'un groupe de chaque condition, avec le groupe le plus important des conditions comportant la séquence des variants C8 shuffle, E2 shuffle, Alanine et Sac7d sauvage, mais également le second groupe généré avec les variants C6, C8 et C6 shuffle. Le second pôle qui semble privilégié comporte quant à lui le groupe le plus important des conditions issues des variants C6, C8, C6 shuffle, ainsi que le second groupe des conditions E2, C8 shuffle et E2 shuffle. Forts de ces constatations, nous avons conclu que les deux régions de la GFP ciblées par ces pôles préférentiels présentaient un fort potentiel comme zones d'intérêt pour le design d'interface protéine-protéine.

Dans un second temps, nous avons reporté ces meilleurs résultats d'ancrage moléculaire à la surface de la GFP en identifiant individuellement les résidus ciblés. Moins simplistes, les représentations ainsi obtenues sont tout de même en corrélation avec les groupes obtenus en

représentations en centroïdes. En considérant chaque condition séparément (**Figure 58**), il est possible d'observer que les régions hautes et basses du tonneau beta de la GFP sont rarement ciblées, tout comme une de ses faces planes. Les autres faces, situées autour des centroïdes des groupes préalablement identifiés, coïncident avec les deux pôles identifiés tout en apportant une information supplémentaire concernant l'étendue de ces régions.

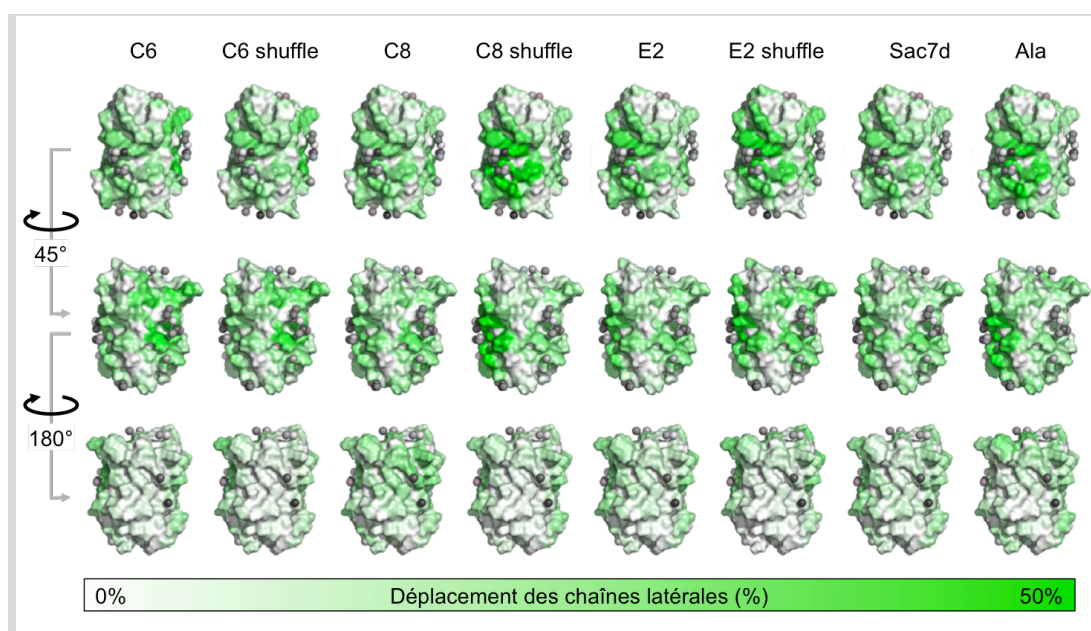


Figure 58: Représentation des surfaces de la GFP les plus fréquemment ciblées. Les acides aminés modifiés dans les simulations d'ancrage moléculaire (chaîne latérale décalée) sont représentés en gradient de vert, proportionnellement à leur fréquence d'implication dans les complexes Nanofitine:GFP prédits.

Par ailleurs, ces observations sont d'autant plus évidentes en considérant la totalité des meilleures poses d'ancrage moléculaire (**Figure 59A**). Les régions figurées en vert à la surface de la GFP pourraient donc représenter des surfaces propices à la formation d'interfaces protéine-protéine d'après les simulations effectuées. Ce résultat a été conforté lors de la comparaison avec les co-cristaux résolus de Nanobodies modulateurs de fluorescence, dirigés spécifiquement contre la GFP avec des affinités de l'ordre du nanomolaire, en complexe avec leur cible (identifiants PDB: 3K1K et 3G9A; Kubala *et al.*, 2010). Les surfaces colorées en vert et en bleu sur la **Figure 59B** correspondent, respectivement, aux interfaces avec le Nanobody capable d'induire une diminution de l'émission de fluorescence de la GFP et avec celui capable d'induire son augmentation. En comparant ces interfaces caractérisées expérimentalement avec nos prédictions, nous avons constaté que le premier pôle prédit chevaucherait l'interface avec le

Nanobody réducteur de fluorescence. D'autre part, le second pôle prédit chevaucherait l'interface formée avec le Nanobody amplificateur de fluorescence, et plus particulièrement au niveau d'une région hydrophobe observée à sa périphérie. Ceci soulignerait donc une corrélation importante, via une validation expérimentale indirecte, entre la prédiction de surfaces à cibler sur la GFP et la possibilité de diriger effectivement des protéines d'affinité à leur rencontre.

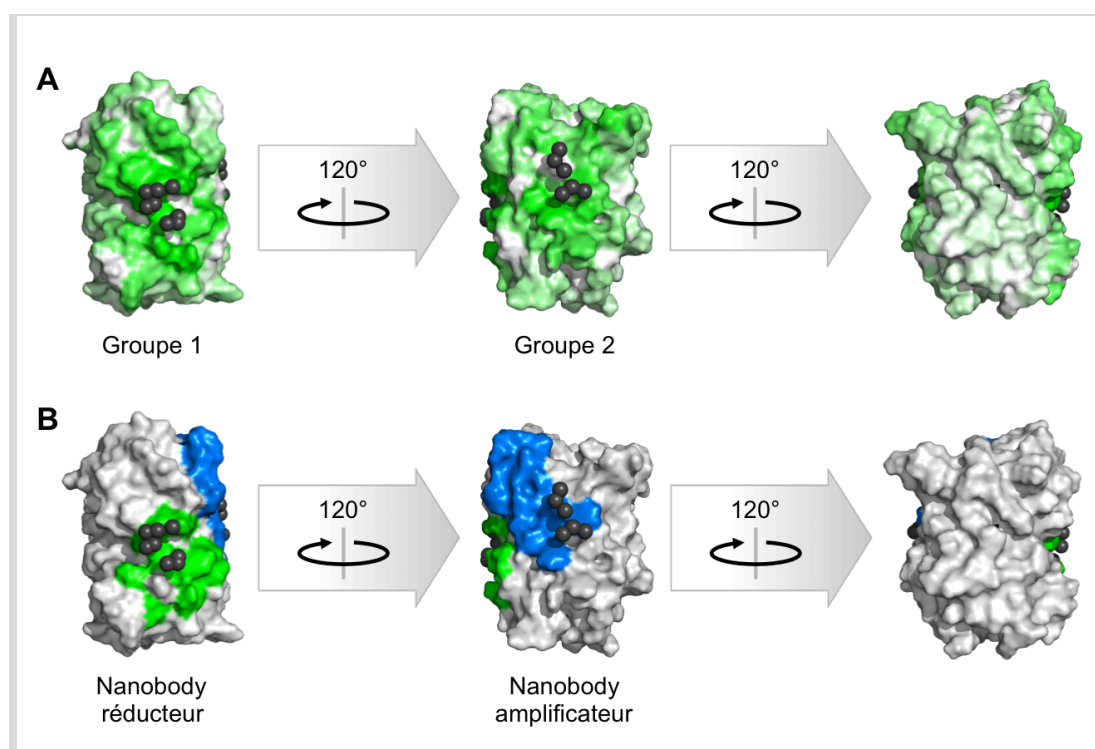


Figure 59: Représentation des surfaces de la GFP favorables à des interactions de type protéine-protéine. Les centres des groupes préalablement identifiés sont représentés par des sphères (en noir) à la surface de la GFP (en gris). A) Les acides aminés modifiés dans les simulations d'ancrage moléculaire (chaîne latérale décalée) sont représentés en gradient de vert, proportionnellement à leur fréquence d'implication dans les complexes Nanofitine:GFP prédits. B) Les surfaces d'interaction avec les Nanobodies réducteur et amplificateur de fluorescence sont représentées, respectivement, en vert et en bleu.

Alors que les expériences d'ancrage moléculaire n'ont pas permis de ségrégation efficace entre des Nanofitines dirigées contre la GFP et des Nanofitines non spécifiques, elles sembleraient donc plus fiables en ce qui concerne l'identification de régions à fixer à la surface de la GFP. Si ce type de résultat s'avérait reproductible sur des cibles de nature différente, cet aspect prédictif pourrait jouer un rôle encore plus important, notamment dans des cas de prédictions d'interface sur des cibles sans information expérimentale disponible. Cette méthode pourrait également se révéler extrêmement utile pour évaluer la possibilité de cibler un domaine protéique spécifique à l'aide

de Nanofitines, y compris dans le cas de récepteurs aux petites molécules pour lesquels des interfaces favorables aux interactions protéine-protéine ne sont pas nécessairement présentes.

III.3.4. Orientation des Nanofitines pour la formation d'interfaces

En plus d'identifier les deux régions à la surface de la GFP qui semblent favorables à la formation d'interfaces avec d'autres protéines, nous avons cherché à caractériser les surfaces de contact prédites entre les Nanofitines et la GFP lors des simulations d'ancrage moléculaire, avant de modifier leur composition par design protéique. Pour cela, nous avons tout d'abord calculé les surfaces accessibles de différents éléments: chaque complexe formé dans sa totalité, la GFP ou la Nanofitine dissociée de son partenaire, ainsi que les résidus situés dans le second feuillet beta de la Nanofitine. Grâce à cela, nous avons estimé la surface d'interaction totale ainsi que la contribution du feuillet beta de la Nanofitine pour chaque complexe, comme schématisé sur la **Figure 60**. Les surfaces d'interaction ainsi calculées et combinées aux différents scores fournis par Rosetta, en particulier ceux d'énergie totale ou d'interaction, ont permis d'énoncer des règles générales quant à la prédiction d'interface de type protéine-protéine.

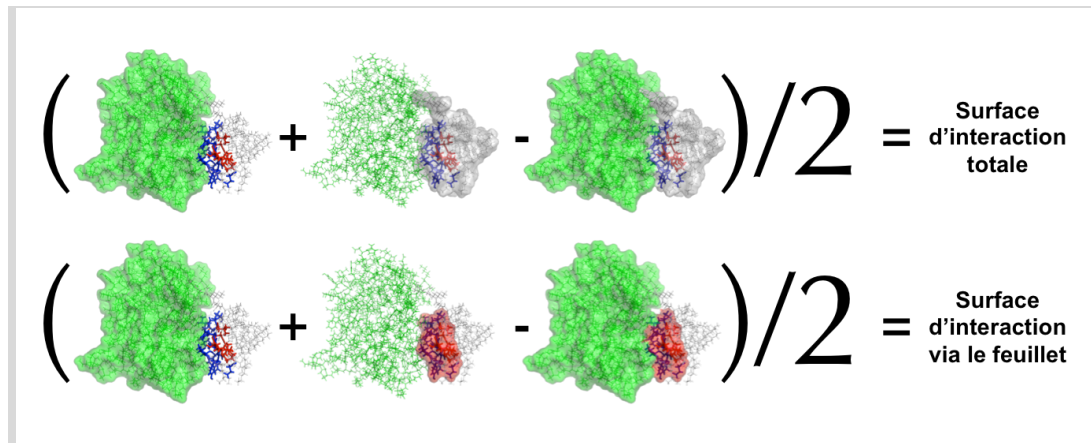


Figure 60: Les atomes des structures sont représentés en lignes, de couleur verte pour la GFP, et grise pour la Nanofitine à l'exception des résidus du second feuillet beta (en bleu ou en rouge, respectivement, s'il sont exposés ou enfouis). La surface accessibles prises en compte dans les calculs sont représentées en vert, gris et rouge, respectivement, pour la GFP, la Nanofitine entière et le second feuillet beta de la Nanofitine. Les surfaces accessibles, calculées avec une sonde de 1,4 Å de rayon par SurfaceRacer, permettent de définir les surfaces d'interactions d'après les relations:

$$Interface\ totale = \frac{GFP + Nanofitine - Complexe}{2} \text{ ou } Interface\ via\ feuillet = \frac{GFP + Feuillet - Complexe}{2}$$

Tout d'abord, la distribution des scores d'énergie d'interaction (**Figure 61A**) montre que la plupart des simulations d'ancrage moléculaire ont généré des complexes avec des scores supérieurs à -4, aboutissant en moyenne à 246 complexes satisfaisants compris entre -5 et -10, par séquence de

Nanofitine évaluée. Cette proportion permet d'estimer l'ampleur de l'espace conformationnel à explorer dans ce type de simulations, générant principalement des complexes de moindre intérêt avant d'aboutir à des solutions jugées satisfaisantes. Nous pouvons également noter que les huit conditions menées en parallèle ont finalement mené à des profils similaires de scores d'interaction, confortant l'hypothèse selon laquelle la géométrie des Nanofitines a eu un impact plus important que celui des résidus exposés à leur surface dans ces prédictions. Un impact des séquences a pourtant été observé lorsque nous nous sommes focalisés sur les énergies totales calculées par Rosetta (**Figure 61B**), indiquant notamment des profils moins stables avec les Nanofitines E2 ou E2 shuffle. Comme préalablement discuté dans la section "II.5.2.2. Stabilité" du chapitre dédié à l'humanisation de Nanofitine, ces conditions arborent des énergies moins favorables en raison de la présence d'une proline potentiellement déstabilisante dans la séquence de la Nanofitine anti-GFP E2. Néanmoins, nous avons pu confirmer une meilleure stabilité des complexes présentant des scores d'interaction compris entre -5 et -10 dans chaque condition, avec un décalage vers des scores totaux d'énergie plus négatifs.

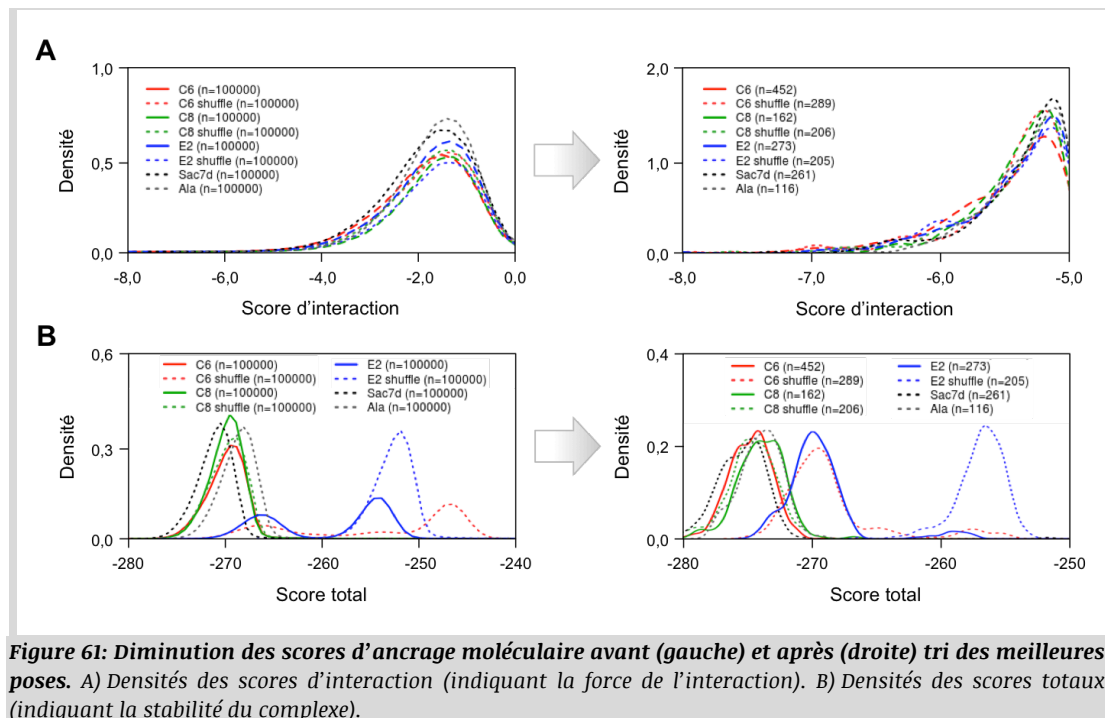


Figure 61: Diminution des scores d'ancrage moléculaire avant (gauche) et après (droite) tri des meilleures poses. A) Densités des scores d'interaction (indiquant la force de l'interaction). B) Densités des scores totaux (indiquant la stabilité du complexe).

Au delà de ces scores d'énergie, le calcul des surfaces masquées lors de la formation des complexes (**Figure 62A**) a montré que le processus d'ancrage moléculaire a généré une majorité de complexes avec des surfaces d'interactions comprises entre 0 et 400 Å². Or, il est généralement décrit que la formation d'interfaces non covalentes entre protéines fait suite à l'enfouissement de surfaces comprises entre 300 et plus d'un millier d'Ångström² (Janin, 1997; Chen et al., 2013; Baskaran et al., 2014). De façon relativement intuitive, il a également été observé une corrélation entre l'augmentation d'affinité (c'est-à-dire une diminution de K_D) et l'augmentation de la surface enfouie lors d'une interaction (Chen et al., 2013). Ce phénomène semble avoir été simulé lors des prédictions d'ancrage moléculaire, puisque les scores d'interactions calculés sont plus favorables à mesure que la surface d'interaction augmente (**Figure 62B**), aboutissant aux meilleurs résultats présentant des surfaces enfouies principalement comprises entre 300 et 700 Å².

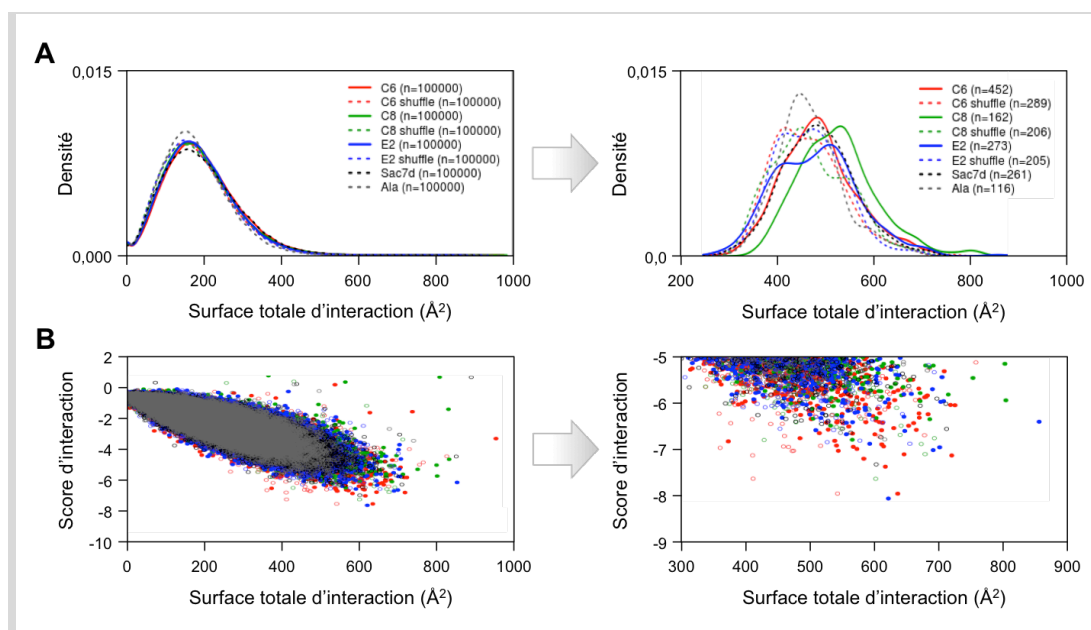


Figure 62: Augmentation des surfaces d'interaction obtenues en ancrage moléculaire avant (gauche) et après (droite) tri des meilleures poses. A) Densités des surfaces totales d'interaction. B) Nuages de points des scores d'interaction, représentés en fonction des surfaces totales d'interaction des complexes prédits.

Afin de projeter ces résultats à l'échelle de la charpente des Nanofitines, nous pouvons noter que Sac7d enfouit environ 16,8% de sa surface exposée au solvant lors de la formation d'un complexe avec de l'ADN sous forme double brin (dans la structure 1AZP), ce qui représente 833 Å² de la surface de la protéine. A cela, il est important d'ajouter la variabilité qui peut entrer en jeu en

fonction de la longueur et de l'organisation spatiale des groupements fonctionnels exposés en surface de chaque Nanofitine. Par exemple, muter les résidus de Sac7d pour générer les Nanofitines anti-GFP C6, E2 ou C8 représente respectivement une modification de surface de +297, +41 ou -12 Å², permettant d'accéder à des affinités nanomolaires.

Par conséquent, ces résultats d'ancrage moléculaire signifient qu'on peut s'attendre à ce qu'une majorité des solutions proposées ne soient pas représentatives d'interfaces biologiquement possibles, comme en témoignait la faible proportion de simulations aboutissant à des scores d'interactions entre -5 et -10. Néanmoins, les meilleurs résultats prédits semblent compatibles avec la formation d'interface protéine-protéine et se rapprochent de la taille de la surface naturellement présente sur Sac7d pour se lier à son ligand.

Par ailleurs, la surface d'interaction naturelle de Sac7d est essentiellement située sur son second feuillet beta. Nous avons également observé l'implication significative de ce feuillet beta lors des expériences d'ancrage moléculaire avec une surface du feuillet enfouie représentant moins de 20 Å² pour la plupart des poses générées, alors que cette surface est augmentée entre 100 et 500 Å² pour les complexes avec les meilleurs scores d'interaction (**Figure 63A**).

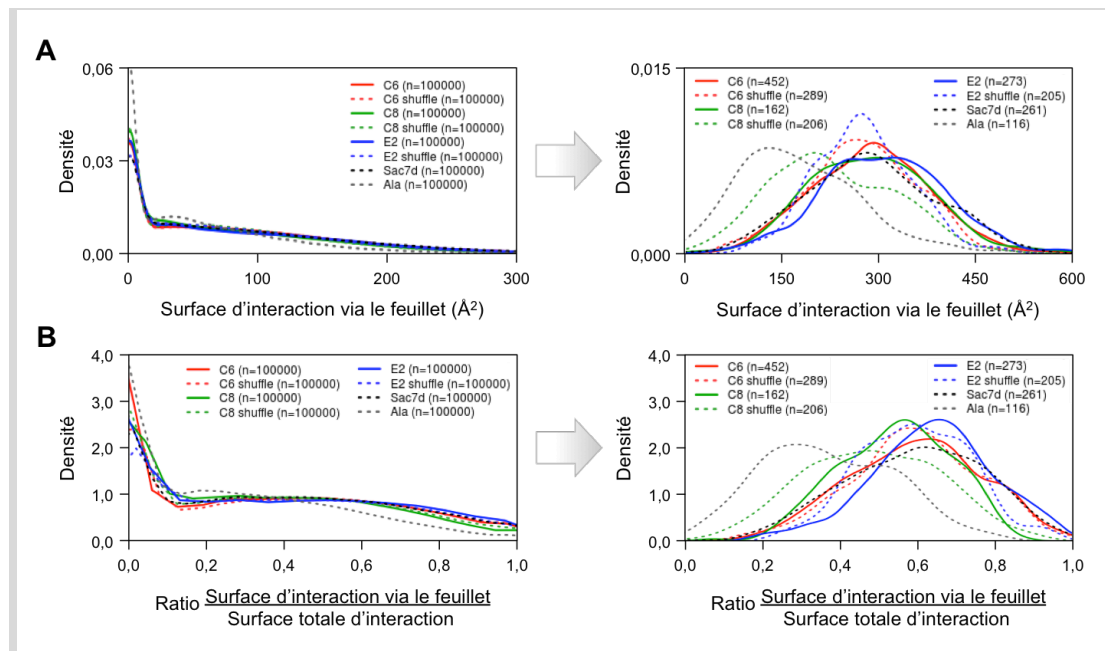


Figure 63: Contribution du feuillet des Nanofitines dans les surfaces d'interaction obtenues en ancrage moléculaire avant (gauche) et après (droite) tri des meilleures poses. A) Densités des surfaces d'interaction via le feuillet. B) Densités des ratio de la surface d'interaction via le feuillet par rapport à la surface totale d'interaction.

D'une part, cette orientation préférentielle confirme qu'une liberté suffisante a été autorisée lors des simulations, laissant la possibilité aux Nanofitines de fixer la GFP par d'autres interfaces que leur feuillet beta. En règle générale, l'inspection manuelle de résultats parmi les moins satisfaisants de ceux proposés par Rosetta a effectivement mis en évidence de petites surfaces d'interaction ou des interactions via l'hélice de Sac7d. D'autre part, ceci met en avant une contribution des résidus du feuillet de Sac7d dans la fixation à la GFP dans les meilleures poses, bien que cela ne représente pas l'intégralité de l'interface formée à la surface de la Nanofitine (**Figure 63B**). Ceci implique donc que des résidus qui n'auraient pas été mutés aléatoirement par sélection *in vitro*, comme avec la banque utilisée pour générer les Nanofitines anti-GFP à humaniser décrites précédemment, puissent être intégrés à l'interface formée avec la GFP. Ce type de résultat pourrait donc présenter un apport pour le design rationnel de banques à utiliser en *ribosome display*, en indiquant les résidus dont la mutation pourrait générer des protéines de forte affinité. Il serait également intéressant de reproduire ce type d'analyse sur une cible avec une surface moins plane que celle du tonneau beta de la GFP, pour déterminer si l'orientation des Nanofitines restait en faveur d'une interface impliquant leur second feuillet beta. De plus, nous savons que des protéines spécifiques de la GFP peuvent s'y lier grâce à des domaines variables localisés dans leurs boucles, comme nous l'avons notamment observé avec les Nanobodies anti-GFP dont les épitopes chevauchent les sites découverts par ancrage moléculaire.

Plus généralement, l'implication du feuillet des Nanofitines dans ces simulations renforce l'intérêt de muter la surface naturelle de liaison à l'ADN de Sac7d dans le cadre de la génération de Nanofitines anti-GFP. En plus de présenter une complémentarité géométrique avec la protéine ciblée, l'interaction via un feuillet beta muté peut contribuer à abroger la fixation à l'ADN, qui n'est pas désirée le plus souvent. De ce fait, le design protéique réalisé en aval a été effectué sur les 14 positions décrites dans la première conception de banque (**Figure 27**), mutant ainsi les résidus de Sac7d impliqués dans la liaison à l'ADN (désignée ci-après comme la banque). Au final, nous avons décidé de sélectionner les poses d'ancrage moléculaire présentant les meilleurs scores d'interaction pour initier le design protéique. Indirectement, ceci correspond donc aux

orientations favorisant une fixation via le feuillet des Nanofitines et stabilisant les complexes par une augmentation de leur surface d'interaction.

III.3.5. Sélection des complexes pour le design protéique

Après avoir généré et caractérisé des complexes issus des prédictions d'ancrage moléculaire, nous avons sélectionné des poses de départ pour chacun des pôles favorables aux interactions Nanofitine:GFP identifiés à la surface de la GFP. Pour cela, nous avons conservé les complexes possédant des scores d'interaction compris entre -5 et -10, et dont au moins 5 carbones alpha de la banque se situent à moins de 10 Å du centre d'un des pôles d'interaction. Grâce à cette sélection, nous avons limité le nombre de conditions à engager en design protéique en conservant les poses les plus proches des pôles à cibler, tout en permettant aux Nanofitines d'avoir des orientations suffisamment différentes. Ceci représente 27 et 22 poses de départ pour le design *in silico*, respectivement pour le premier et le second pôle (**Figure 64**).

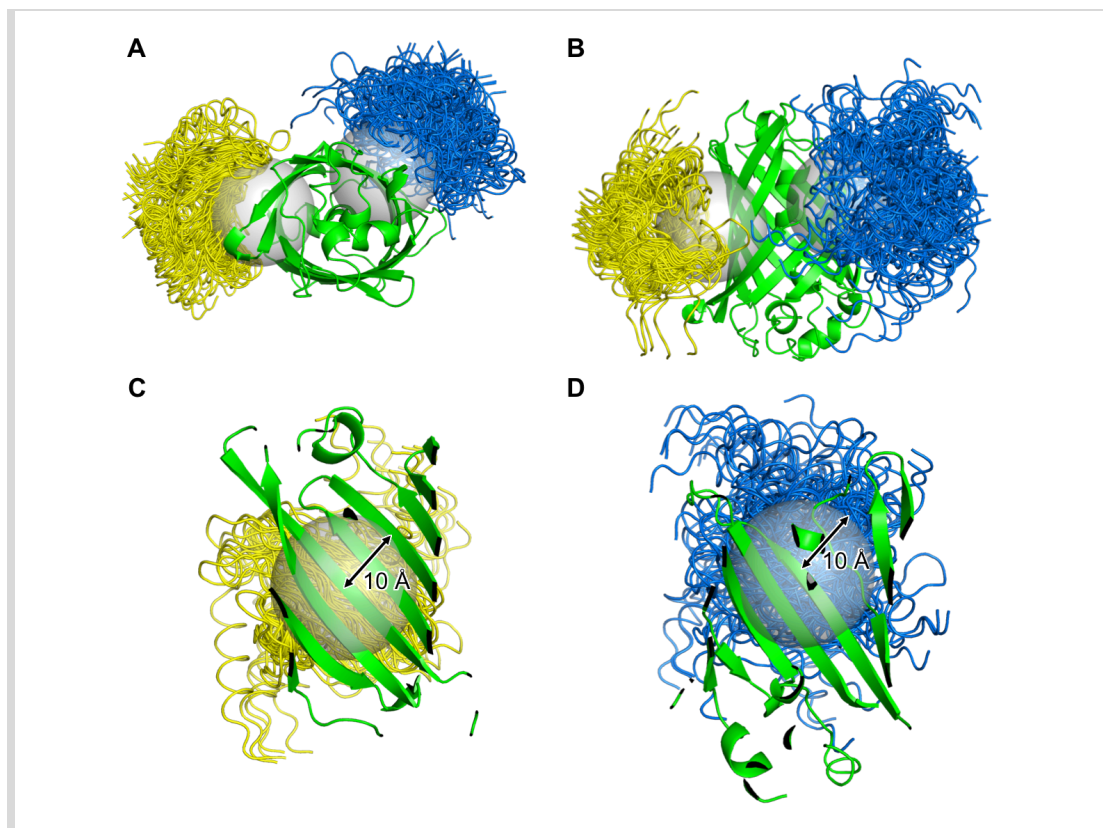


Figure 64: Poses des complexes identifiés par ancrage moléculaire pour le design *in silico* de Nanofitines. Une sphère de 10 Å de rayon (en gris) a été définie autour du centre de chacun des pôles identifiés à la surface de la GFP (en vert). Les structures de Nanofitines sélectionnées, dont au moins 5 carbones alpha des résidus de la banque sont situés dans une des sphères, sont représentées en rubans de couleur jaune (groupe 1) ou bleue (groupe 2). A) Vue de dessous. B) Vue de face. C) Vue du groupe 1 (constitué de 27 poses) depuis l'intérieur de la GFP. D) Vue du groupe 2 (constitué de 22 poses) depuis l'intérieur de la GFP.

En considérant toutes les étapes de prédictions d'ancrage moléculaire et de traitement de ses résultats, nous avons identifié les 49 complexes qui nous ont parus les plus prometteurs à partir d'un total de 800 000 complexes générés. Ceci représente une infime partie de l'espace effectivement exploré lors des simulations. Cette exploration bénéficie des performances améliorées des calculs de Rosetta, notamment via l'utilisation de matrices optimisées pour l'exploration des rotamères ou pour les calculs de scores entre paires d'acides aminés. Le temps de calcul nécessaire devrait également devenir plus accessible et pouvoir être réduit en matière de "temps humain", en accord avec les conjectures dérivées de la loi de Moore (*Moore, 1965*) qui annoncent une augmentation constante de la puissance des ordinateurs.

Malgré tout, l'exploration d'un tel espace pourrait bénéficier d'optimisations supplémentaires, par exemple en croisant les sources d'informations (prédictives ou observées expérimentalement) pour être concentrée sur des régions mieux définies à la surface de la cible à lier. L'exploration de l'intégralité de la surface de la GFP, telle que présentée dans ces travaux, représente d'ailleurs un challenge additionnel qui a pour principale vocation de mettre à l'épreuve les méthodes de calcul de Rosetta dans le cadre d'une preuve de concept. Pour des applications futures, le design rationnel de Nanofitines n'apportera une plus-value que s'il accompagne la génération de protéines affines et spécifiques d'un épitope défini au préalable. En effet, contrôler cette spécificité à la surface de leur cible permettrait dans de nombreux cas de prédire indirectement la fonction des Nanofitines, en particulier lors de la génération d'agents neutralisants. Dans la preuve de concept présentée dans ce chapitre, une situation comparable pourrait par exemple être représentée par les interactions Nanobody:GFP à mimer ou à inhiber, ce qui aurait eu pour effet de nous faire converger plus rapidement vers les deux pôles de la GFP que nous avons identifié *de novo*.

III.4. Design protéique: Optimisation de l'interface par ingénierie de la surface de la Nanofitine

III.4.1. Description de l'algorithme

Nous avons identifié par ancrage moléculaire 49 orientations de complexes comme poses les plus favorables à la formation d'interface entre Nanofitine et GFP. De plus, les régions ciblées nous sont apparues en adéquation avec la littérature tant par comparaison avec des complexes connus que par l'étendue des surfaces concernées. L'interface Nanofitine:GFP a ensuite été modifiée par design de la surface de la Nanofitine à l'aide de l'application *fixbb* de Rosetta, afin de prédire des Nanofitines affines pour la GFP et spécifiques des épitopes visés.

Effectuer ce design protéique *in silico* correspond à répondre à un problème d'optimisation combinatoire pour identifier les meilleurs groupements fonctionnels et leur conformation spatiale à exposer à la surface des squelettes peptidiques fixes pour stabiliser l'interaction. De fait, ceci représente un problème NP-complet dont il est difficile de trouver efficacement des solutions, bien qu'il soit possible de vérifier rapidement les solutions proposées (*Pierce et Winfree, 2002*). L'algorithme mis à profit par Rosetta ne peut donc pas garantir que les chaînes latérales proposées constituent les meilleures solutions, mais les temps de calcul nécessaires sont considérablement réduits grâce à son approche stochastique de recuit simulé (*Kirkpatrick, 1984; Černý, 1985*) évitant le calcul préalable des énergies de toutes les paires de rotamères (*Leaver-Fay et al., 2005, 2008*). Dans cette étude, nous avons contraint Rosetta à tester les rotamères de la GFP et de la Nanofitine sans introduire de mutations en dehors des 14 positions de la banque de Nanofitines. Nous avons autorisé l'exploration de l'ensemble des acides aminés naturels à ces positions rendues aléatoires, à l'exception des résidus prolines et cystéines pour limiter, dans un cas, les risques de modification de la conformation de la Nanofitine et, dans l'autre, la formation de ponts disulfures masquant une partie du site d'interaction ou favorisant l'agrégation.

Avant de discuter des résultats après obtention de 1000 solutions de design protéique par pose de départ, nous devons souligner les limitations importantes de l'approche que nous avons mise en place. Premièrement, ces prédictions ont été réalisées avec des squelettes peptidiques fixes.

N'introduisant pas de mutations dans la séquence de la GFP, ceci ne devrait donc pas impacter significativement la conformation de la protéine cible. Cependant, jusqu'à 14 mutations ont été effectuées à la surface des Nanofitines prédites. Bien que l'application aux Nanofitines semble judicieuse puisque leur charpente a déjà démontré sa capacité à supporter l'introduction d'une variabilité aussi importante, il est difficile d'anticiper les éventuels mouvements induits qui ne seront pas pris en compte sans exploration de l'espace conformationnel du squelette peptidique. Deuxièmement, cette méthode propose une approche stochastique pour répondre à un problème NP-complet, ce qui implique que les résultats peuvent varier si le nombre de simulations réalisées est trop réduit et/ou en utilisant des graines aléatoires différentes. Enfin, les fonctions de score utilisées ne pouvant être parfaites (*Das et Baker, 2008*), l'éventualité de variants ne s'exprimant pas ou ne se repliant pas correctement n'est pas à écarter malgré l'obtention de meilleurs scores.

Lors de cette étude, nous avons modifié l'interface de complexes, préalablement obtenus par ancrage moléculaire entre la GFP et la structure de Sac7d, pour aboutir à 49 000 complexes regroupés par pôles à cibler (groupes 1 et 2, respectivement, pour les pôles 1 et 2 définis lors de l'ancrage moléculaire). Chacun de ces complexes a ensuite été trié sur la base de son score d'interaction, représentant la différence entre l'énergie du complexe et la somme des énergies des partenaires non complexés. Nous avons ensuite considéré les complexes présentant les scores d'interactions les plus avantageux, puis étudié leur surface d'interaction ainsi que leurs mutations introduites par le design protéique.

III.4.2. Convergence vers des interfaces Nanofitine:GFP étendues

Les résultats de design proposés par Rosetta indiquent que les deux groupes présentent des répartitions similaires en ce qui concerne leurs scores d'interaction (**Figure 65A**), avec environ deux tiers des complexes calculés entre -10 et -15, et un tiers entre -15 et -20. Ceci semblerait indiquer que les deux pôles ciblés présentent globalement les mêmes potentiels d'interaction protéine-protéine, selon les critères pris en compte dans cette approche. Il est cependant important de noter que les fonctions de score entre l'ancrage moléculaire et le design sur

squelette peptidique fixe ne sont pas identiques, ce qui ne permet donc pas de comparer directement les scores d'interaction entre ces deux étapes de notre méthode.

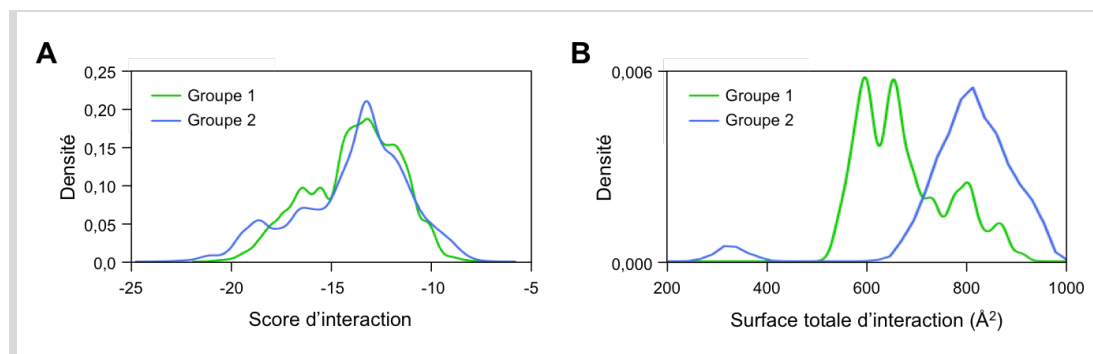


Figure 65: Comparaison des scores (A) et surfaces (B) d'interaction obtenus après design de l'interface des Nanofitines.

Nous avons observé, en accord avec nos attentes, une augmentation de la surface d'interaction des Nanofitines comprise entre 300 et 700 Å² pour les meilleurs résultats d'ancrage moléculaire. Cette extension de l'interface a été accentuée grâce à l'optimisation des chaînes latérales exposées à la surface de Sac7d lors du design, aboutissant à des interfaces comprise entre 500 et 1000 Å² ou entre 600 et 1000 Å², respectivement, pour les groupes 1 et 2 (**Figure 65B**). En comparant ces surfaces, il semblerait également que les deux groupes se distinguent, avec des interfaces formées dans le groupe 2 plus étendues que celles dans le groupe 1. Les distinctions les plus notables ont cependant été observées au niveau des compositions en résidus placés à l'interface par l'algorithme de Rosetta.

En isolant les 200 meilleurs complexes par pôles de la GFP à cibler, nous avons pu observer une convergence spécifique à chaque groupe (**Figure 66**). Peu de mutations communes ont été obtenues entre les deux groupes, ce qui témoigne qu'il ne s'agit pas simplement d'une augmentation de stabilité des Nanofitines indépendante de leur liaison à la GFP, mais bien d'une spécificité des séquences proposées vis-à-vis de l'épitope à fixer. Toutefois, nous avons remarqué dans les deux cas que la plupart des mutations proposées introduisent des acides aminés considérés comme étant plutôt hydrophobes. Plus en détails, les résultats obtenus n'ont pas proposé un résidu unique par position, à l'exception notamment de la mutation K7V observée à la première position variable dans le groupe 2. Pour autant, cette diversité ayant subsisté est en

partie liée à des homologies de groupements fonctionnels (au niveau de leur polarité, hydrophobie ou taille), notamment avec des associations entre résidus lysine ou arginine, sérine ou thréonine, acide glutamique ou acide aspartique.

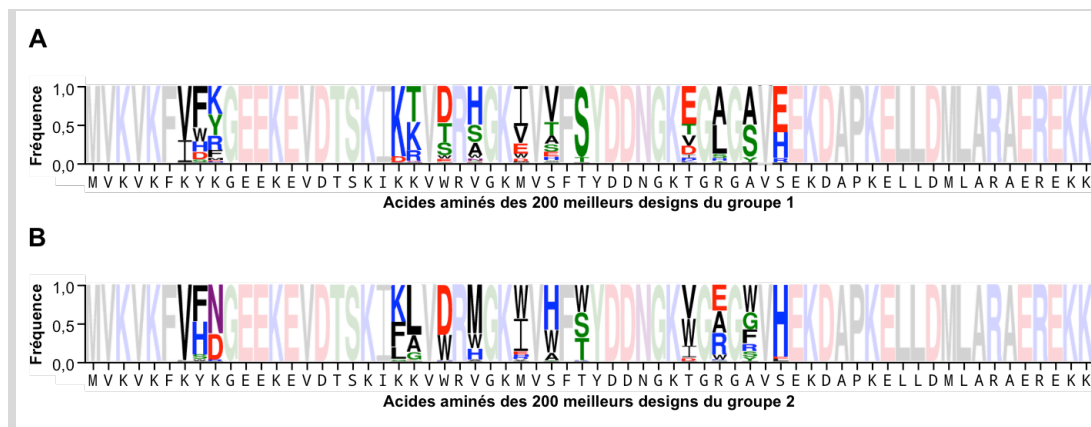


Figure 66: Diversité des séquences des meilleures poses obtenues après design de l'interface des Nanofitines. Représentations des fréquences de mutations introduites dans les meilleurs scores du groupe 1 (A) et du groupe 2 (B), générées par WebLogo (Crooks et al., 2004).

De façon intéressante, la convergence de séquences observée provient d'un nombre limité de poses de départ utilisées pour le design *in silico*. Par exemple, ne considérer que les meilleurs scores jusqu'à obtenir 10% des séquences proposées par Rosetta revient à exclure complètement 19/27 complexes issus d'ancrage moléculaire pour le groupe 1, et 17/22 pour le groupe 2 (**Figure 67**). Cette observation, couplée à la redondance et à la convergence de séquences observée dans les résultats de design, pourrait également signifier que la diversité de variants proposés par notre méthode de design est fortement dépendante des poses de départ employées, puisque le design en lui-même converge rapidement.

Au moment de l'obtention de l'ensemble de ces prédictions, le nombre de séquences a été jugé suffisamment restreint pour construire et exprimer les meilleurs variants. Nous avons donc identifié 10 Nanofitines à produire pour vérifier leur capacité de fixation à la GFP.

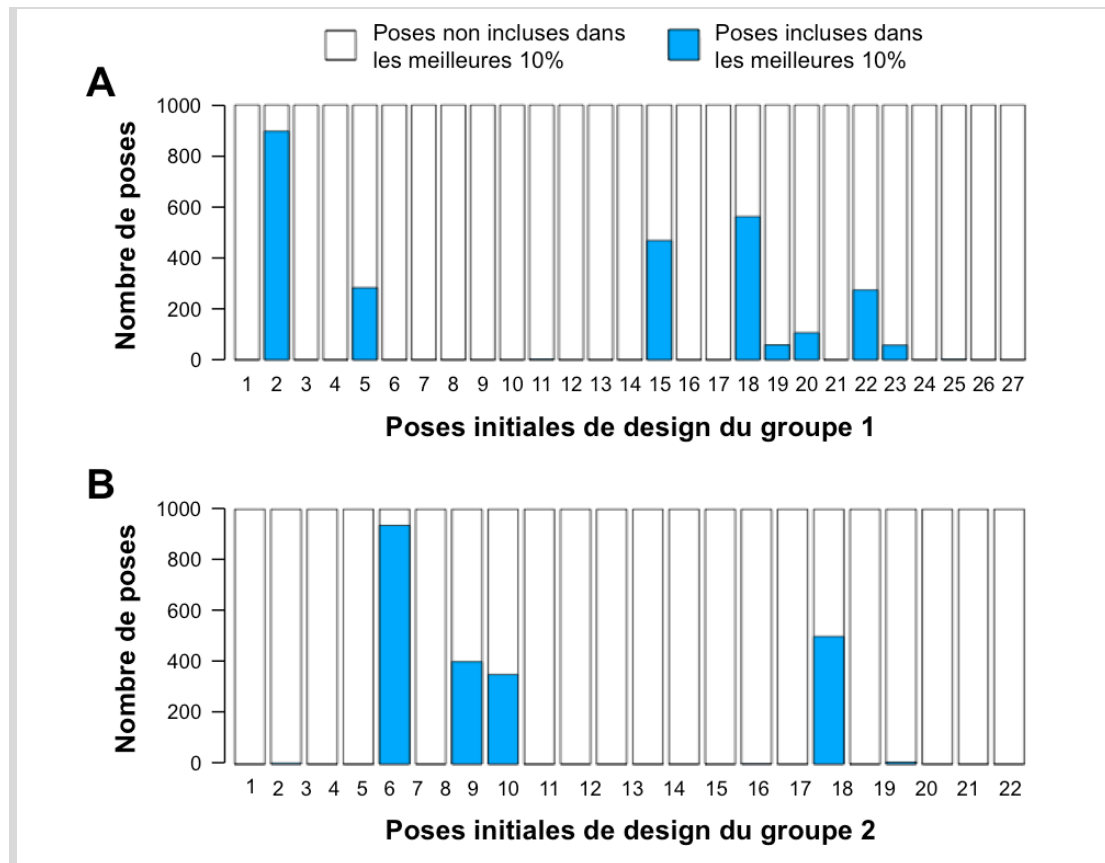


Figure 67: Dépendance des meilleurs résultats de design en fonction de leur pose initiale. Histogrammes en barre pour le groupe 1 (A) et le groupe 2 (B). Les poses avec les scores les moins avantageux (blanc) sont superposées à celles ayant les meilleurs scores (bleu).

III.4.3. Obtention de 10 potentielles Nanofitines anti-GFP

Parmi les meilleurs complexes proposés par design *in silico*, nous avons sélectionné 5 Nanofitines par groupe afin de les exprimer et en purifier plusieurs milligrammes (**Figure 68**). D'un côté, les Nanofitines NF1 à NF5 ciblent le premier pôle à la surface de la GFP et partagent 11 résidus communs sur les 14 rendus variables. De l'autre, deux familles de Nanofitines ont été choisies pour le second pôle de la GFP à fixer, représentées par les Nanofitines NF6 à NF8 ainsi que les Nanofitines NF9 et NF10, partageant respectivement 9 et 13 résidus. Ces séquences illustrent au final trois grandes familles de variants avec les meilleurs scores prédits, dont les motifs n'ont jamais été identifiés au préalable lors du séquençage de Nanofitines anti-GFP générées par *ribosome display*.

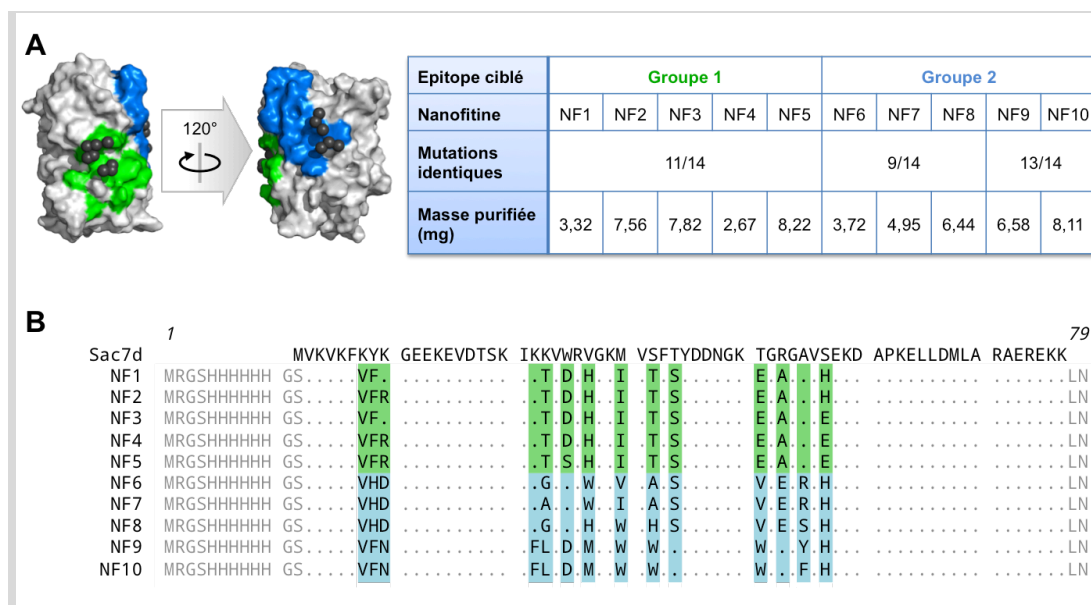
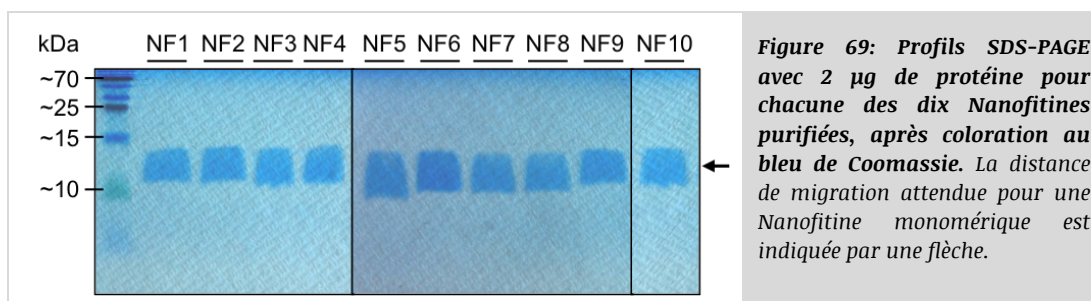


Figure 68: Sélection de 10 potentielles Nanofitines anti-GFP générées in silico. A) Informations résumées des Nanofitines NF1 à NF5, ciblant le premier pôle (vert), et Nanofitines NF6 à NF10, ciblant le second pôle (bleu). B) Alignement des séquences des Nanofitines produites, NF1 à NF10, par rapport à Sac7d. Les positions rendues aléatoires lors du design sont encadrées en couleur.

Les séquences codantes de ces Nanofitines ont été générées par deux réactions successives de PCR, de façon semblable à celle utilisée pour générer les banques de variants de Sac7d (Mouratou et al., 2007), puis clonées dans le vecteur d'expression pAFG01 entre les sites de restriction BamHI et HindIII. Après validation des constructions par séquençage, les 10 Nanofitines ont été exprimées en *E. coli* puis une purification IMAC a été réalisée, aboutissant à l'obtention de 2,67 à 8,82 mg de protéine en une seule fraction d'élution. L'analyse de ces fractions par SDS-PAGE a révélé une pureté satisfaisante se traduisant par la présence d'une bande unique par échantillon, à une distance de migration correspondant à celle d'une Nanofitine monomérique (bien que légèrement décalée vers des plus hautes masses moléculaires, Figure 69).



Chaque variant ainsi exprimé a été obtenu sous forme purifiée et soluble, y compris les Nanofitines NF9 et NF10 malgré l'exposition de nombreux résidus aromatiques hydrophobes à

leur surface. En plus de permettre la caractérisation de ces 10 Nanofitines, les résultats de production en bactérie soulignent l'extrême robustesse de la charpente de Sac7d. En effet, la compatibilité des mutations introduites n'a pas été testée avant expression de clones isolés, contrairement à ce qui est réalisé indirectement au cours du processus de sélection *in vitro* qui élimine les variants instables au sein de la banque.

III.5. Confirmations expérimentales d'une fixation spécifique à la GFP

III.5.1. Criblage des Nanofitines anti-GFP prédites

La fixation des 10 Nanofitines prédites *in silico* à la GFP a été vérifiée par ELISA et par interférométrie de couche biologique (**Figure 70**). Avec les deux techniques, il n'a pas été possible d'observer de signal d'association entre la GFP et les Nanofitines monomériques aux concentrations testées, se traduisant par des signaux de même intensité que ceux observés avec une Nanofitine non affine pour la GFP. Théoriquement, ces deux techniques peuvent mettre en évidence des interactions entre protéines avec des K_D de l'ordre du millimolaire au picomolaire, selon les conditions expérimentales. Cependant, la sensibilité de ces méthodes de détection peut varier, notamment en fonction du k_{off} et de la taille des molécules à détecter, respectivement en ELISA et via interférométrie.

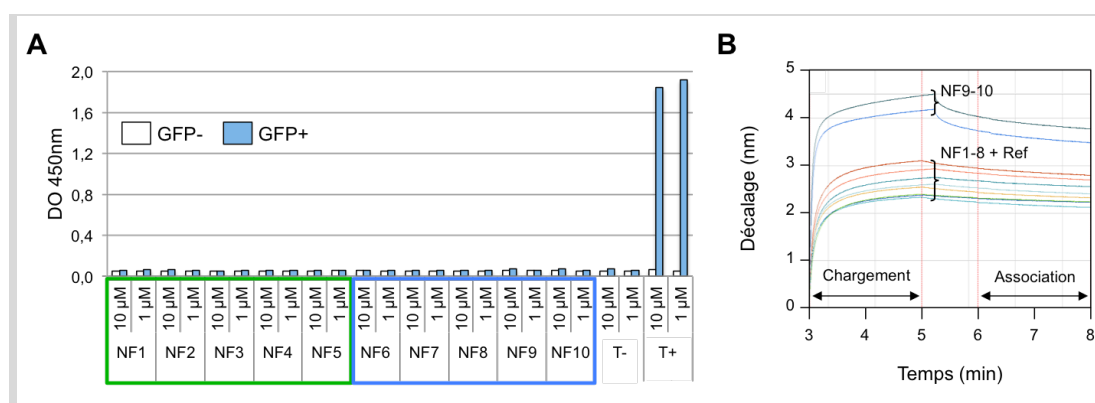


Figure 70: Profils de fixation des Nanofitines purifiées à la GFP, par ELISA et interférométrie. A) ELISA avec 1 ou 10 μM de Nanofitine en présence (barres bleues) ou absence (barres blanches) de StrepTagII-GFP immobilisée. Les Nanofitines ciblant le pôle 1 et 2 sont encadrées, en vert et en bleu. T-: Nanofitine non spécifique. T+: Nanofitine anti-GFP D8. B) Profils cinétiques par interférométrie du chargement des Nanofitines sur les biocapteurs, puis d'association entre les Nanofitines capturées et la StrepTagII-GFP. Ref: Nanofitine non spécifique.

Compte tenu de la difficulté estimée pour découvrir des protéines affines uniquement par une méthode bioinformatique, nous avons supposé que les affinités des Nanofitines testées étaient

sans doute très faibles, voire nulles. Afin d'optimiser les chances de détecter un signal spécifique, nous avons opté pour une modification du mode de présentation des Nanofitines en les multimérisant. Les Nanofitines découvertes *in silico* et la Nanofitine non spécifique de la GFP ont donc été biotinylées chimiquement, puis multimérisées par addition de streptavidine. Les complexes ainsi formés en un temps réduit ont pu présenter jusqu'à 4 Nanofitines, pouvant favoriser un phénomène d'avidité absent en cas d'utilisation de Nanofitines monomériques. Grâce à cette multimérisation, nous avons observé des signaux légèrement plus intenses en ELISA (**Figure 71A**). En particulier, nous avons pu remarquer une augmentation du signal en présence de GFP pour la Nanofitine NF5 dans le groupe 1, de manière significative et dépendante de la concentration de Nanofitines. L'ensemble des autres conditions explorées a résulté en des ratios signal spécifique sur bruit de fond moins intenses, laissant tout de même suspecter un signal spécifique pour NF7 dans la première famille du groupe 2 avec une augmentation de signal par 4 en présence de GFP. Enfin, la dernière famille du groupe 2 a présenté des signaux semblant moins spécifiques, avec les signaux les plus intenses en présence de GFP avec NF9, mais également des signaux significatifs en absence de cible.

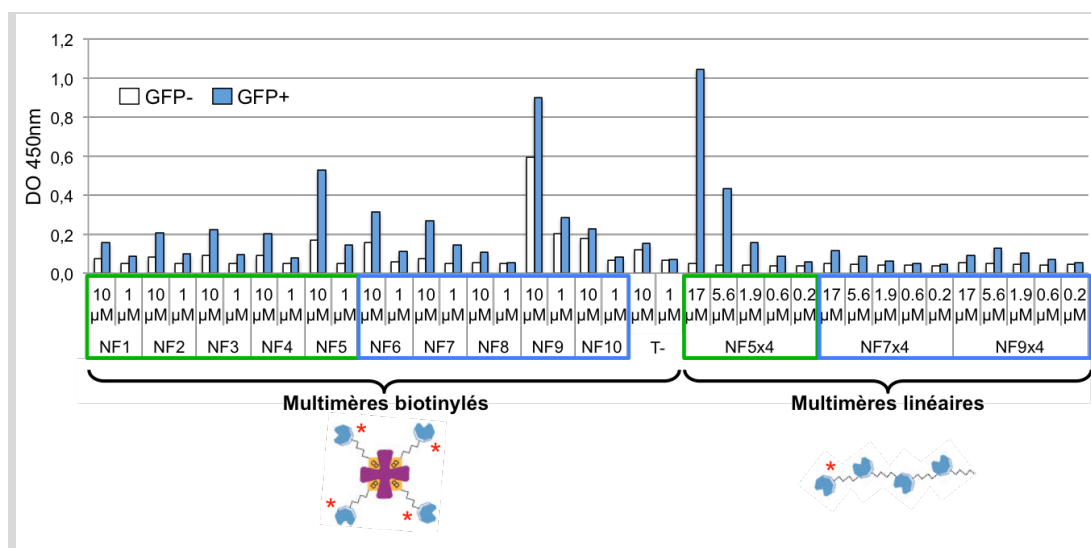


Figure 71: Criblage de fixation à la GFP par différentes concentrations de Nanofitines multimériques, effectué par ELISA. Nanofitines monomériques multiprésentées par couplage biotine/streptavidine (NF1 à NF10) et Nanofitines fusionnées sous forme tétramérique (NF5x4 à NF9x4) révélées par ELISA en présence (barres bleues) ou absence (barres blanches) de StrepTagII-GFP immobilisée. Les Nanofitines ciblant le pôle 1 et 2 sont encadrées, respectivement, en vert et en bleu. T-: Nanofitine non spécifique.

Sur la base de ce criblage effectué par ELISA, nous avons concentré nos efforts sur la caractérisation de multimères linéaires et covalents du variant le plus prometteur par famille. Ce

choix a donc concerné NF5 et NF7, les deux Nanofitines avec le meilleur ratio de signal en présence de cible en comparaison du bruit de fond. La Nanofitine NF9 a également été conservée, bien que les signaux observés en absence de cible peuvent être la conséquence de sa richesse en acides aminés aromatiques hydrophobes et donc d'une fixation non spécifique et/ou de précipitation dans les puits de plaque ELISA. Les résultats obtenus par ELISA à différentes concentrations ont confirmé la fixation spécifique de NF5 à la GFP, tandis que les signaux pour NF7 et NF9 se sont montrés peu intenses et faiblement dépendants de la concentration utilisée (**Figure 71B**).

La différence entre les profils ELISA obtenus lors du criblage par couplage avec biotine et streptavidine et ceux des tétramères linéaires soulève des questions quant à l'efficacité des méthodes de multimérisation. D'une part, la multimérisation par couplage non covalent présente le désavantage de générer une solution potentiellement hétérogène, tant en nombre de sous-unités de Nanofitines multiprésentées que d'orientations du fait du couplage de la biotine sur des amines réactives. Cet inconvénient est potentiellement compensé par des orientations suffisamment variées pour favoriser une présentation multiple efficace, en plus de la rapidité de couplage par rapport à une fusion génétique. D'autre part, les Nanofitines tétramérisées linéairement par fusion génétique présentent des variables inconnues concernant la bonne accessibilité des différentes sous-unités. En effet, il est difficile de déterminer sans étude préliminaire si les bras espaceurs flexibles sont de longueur adaptée, ou si la bonne orientation des sous-unités est assurée (si bien qu'une sous-unité pourrait tout autant être active ou faire office de bras espaceur). De plus, cette méthode de fusion génétique effectuée par assemblage Gibson (*Gibson et al., 2010*) a requis un temps de réalisation plus important, nécessitant à nouveau des amplifications par PCR, un clonage, une validation par séquençage, une production en bactérie, et une nouvelle purification pour chaque construction. Ces incertitudes et désavantages sont néanmoins contrebalancés par l'obtention de plus grandes quantités de tétramères en solutions homogènes comme en témoigne l'analyse SDS-PAGE des multimères linéaires produits et purifiés (**Figure 72**).

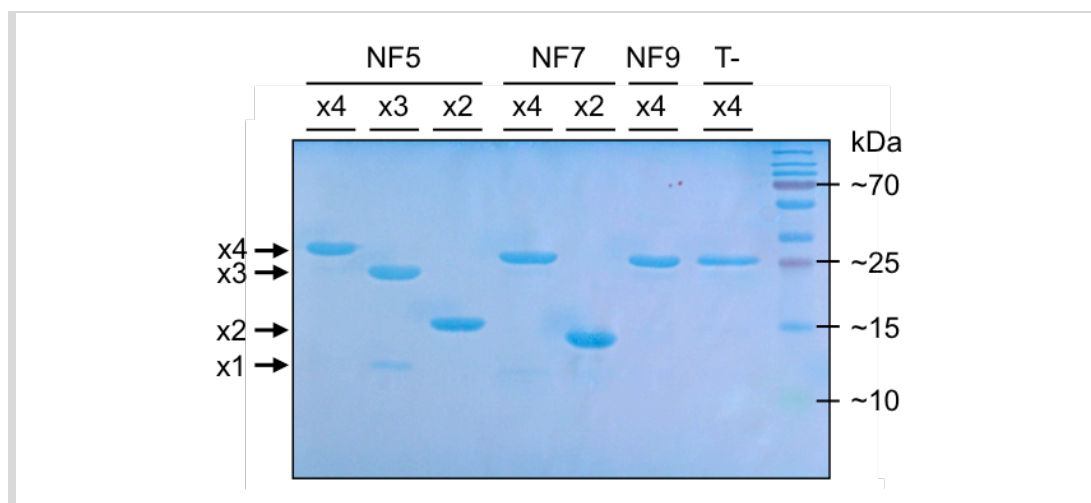


Figure 72: Profils SDS-PAGE avec 2 μ g de protéine pour chacune des Nanofitines multimériques purifiées, après coloration au bleu de Coomassie. Les distances de migration attendues pour une Nanofitine mono-, di-, tri- et tétramérique sont indiquées par une flèche annotée, respectivement, x1, x2, x3 et x4. T- = Nanofitine non spécifique de la GFP.

Indépendamment de ces distinctions méthodologiques, NF5 a été la Nanofitine prédite ayant présenté le meilleur profil de fixation en ELISA. La tétramérisation nécessaire pour visualiser des signaux spécifiques suggère que l'affinité de ce variant envers la GFP est extrêmement modeste (K_D sans doute $\geq \mu$ M). Toutefois, ces résultats semblent reporter que NF5 est spécifique de la GFP, et donc un potentiel succès de notre approche. De ce fait, nous avons concentré la suite des expériences sur l'étude de NF5 sous sa forme tétramérique linéaire pour confirmer ses propriétés.

III.5.2. Validation de l'épitope ciblé par NF5

La Nanofitine NF5 a démontré un profil semblant indiquer sa capacité de liaison à la GFP avec une affinité relativement faible. Si cette spécificité est un point essentiel pour la validation des résultats de prédictions, il nous a néanmoins paru nécessaire de confirmer qu'elle se fixe effectivement au bon épitope, avec une orientation telle que définie *a priori*.

Pour effectuer cette validation, nous avons tout d'abord analysé plus finement les structures de complexes prédites par Rosetta. Notamment, l'analyse des scores d'interaction recalculés par acide aminé ont indiqué une implication prédominante du résidu F99 mais également des résidus K162, D102 et E95 localisés à la surface de la GFP. Ces scores de Rosetta ont ensuite été confirmés lors de la soumission de la pose auprès du service PDBsum, maintenu par l'institut européen de bio-informatique (EMBL-EBI; Laskowski *et al.*, 1997; De Beer *et al.*, 2014). Traitées

comme une résolution de structure cristallographique, les poses prédites mettraient ainsi à profit 61 interactions de type van der Waals, 3 liaisons hydrogène et 1 pont salin pour établir une interface de 894 \AA^2 et 837 \AA^2 à la surface de la Nanofitine et de la GFP, respectivement (**Figure 73**). Toutefois, des distinctions ont subsisté entre l'analyse effectuée via Rosetta et via PDBsum, comme l'implication prédite du résidu K162 qui n'est pas identifiée dans la représentation schématique des interactions.

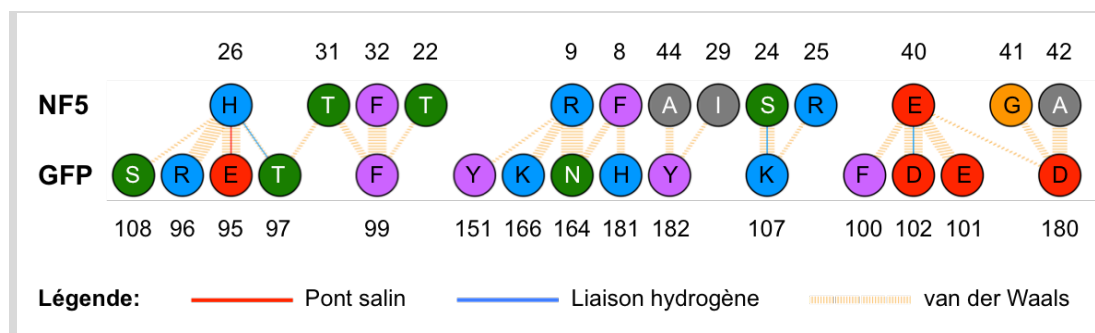


Figure 73: Représentation des interactions prédites par PDBsum entre les résidus de NF5 et de la GFP. Pour les interactions de type van der Waals, l'épaisseur de la bande hachurée est proportionnelle au nombre de contacts atomiques. Les acides aminés sont colorés selon les propriétés de leur groupement fonctionnel, en bleu (Positif: H,K,R), rouge (Négatif: D,E), vert (Neutre: S,T,N,Q), gris (Aliphatique: A,V,L,I,M), mauve (Aromatique: F,Y,W), orange (Proline et glycine: P,G), ou jaune (Cystéine: C).

La confirmation idéale de nos prédictions devant valider à la fois l'épitope ciblé et l'orientation de la Nanofitine à la surface de la GFP, nous avons initié un criblage de conditions de cristallographie du complexe. Pour cela, nous avons exprimé NF5 ainsi que la GFP sans son extrémité C-terminale labile en *E. coli*, avant de les purifier par IMAC et par chromatographie d'exclusion stérique (SEC) pour finalement les placer dans un tampon TBS 0,5X. Compte tenu de la faible affinité attendue pour l'interaction NF5:GFP, nous avons effectué le même traitement avec des protéines chimériques entre NF5 et la GFP, construites par biologie moléculaire de façon similaire aux tétramères de NF5. Pour ces protéines de fusions, chaque domaine a été espacé par un bras flexible de 7, 10 ou 20 résidus (composé de glycines, sérines et alanines), pour permettre un compromis entre la distance nécessaire à la fixation entre NF5 et la GFP et la labilité pouvant inhiber la formation de cristaux. Ainsi, quatre types de complexes ont été préparés (**Figure 74A**) dans l'espoir de pouvoir obtenir des cristaux exhibant une interaction intra- ou intermoléculaire entre NF5 et la GFP. Après la dernière étape de purification, 80 à 100 mg de protéine ont été envoyés par condition à l'équipe de Leonardo Scapozza (Université de Genève, Suisse), avec une

pureté estimée suffisante d'après l'analyse réalisée par SDS-PAGE (**Figure 74B**) qui révèle une bande intense aux tailles attendues pour chaque protéine recombinante.

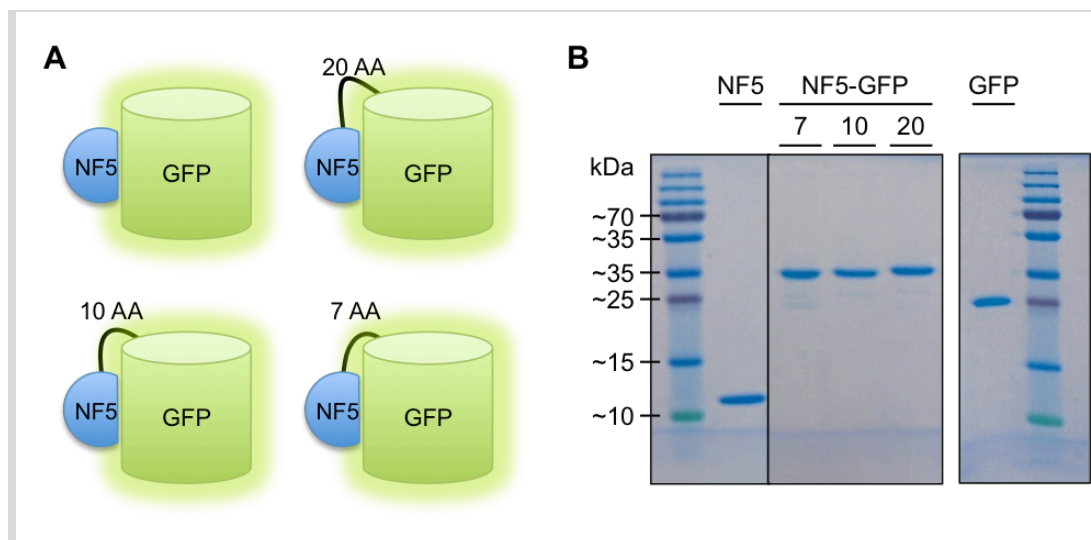


Figure 74: Constructions impliquant NF5 et la GFP pour la cristallographie aux rayons X du complexe. A) Représentation schématique des complexes à cristalliser, non covalent ou fusionnés par un espaceur flexible de 7, 10 ou 20 résidus. B) Profils SDS-PAGE avec 2 µg de protéine pour chacune des protéines purifiées par IMAC et par SEC, après coloration au bleu de Coomassie. Les masses moléculaires des constructions sont attendues à 8,97 kDa (NF5), 35,94 kDa (NF5-7-GFP), 36,12 kDa (NF5-10-GFP), 36,83 kDa (NF5-20-GFP) et 28,50 kDa (GFP).

Nous avons ensuite mené un criblage des conditions de cristallogénèse à partir des protéines chimériques comportant les bras espaceurs flexibles de 7 et 20 résidus, avec l'aide d'Andreja Vujicic-Zagar et de Magali Zeisser-Labouèbe. En résumé, 384 solutions commerciales composées de divers précipitants ainsi que 3 ratios protéine:précipitant par solution ont été explorés par la méthode de goutte assise (ou *sitting-drop*) pour identifier des conditions favorables à la formation de cristaux. Malheureusement, l'observation des gouttes conservées à température ambiante n'a pas permis de déceler de formation de cristaux après un peu plus d'un mois (**Figure 75**). Une des pistes avancées pourrait être la solubilité trop élevée de NF5 et/ou de la GFP, rendant moins accessible les phases de précipitation ou de cristallogénèse (Asherie, 2004). De plus, l'ensemble des protéines purifiées pour la cristallographie a pu être concentré entre 71 et 103 mg/ml (correspondant à des concentrations molaires de 2 à 8 mM) sans observer de précipitation, ce qui conforterait cette hypothèse.

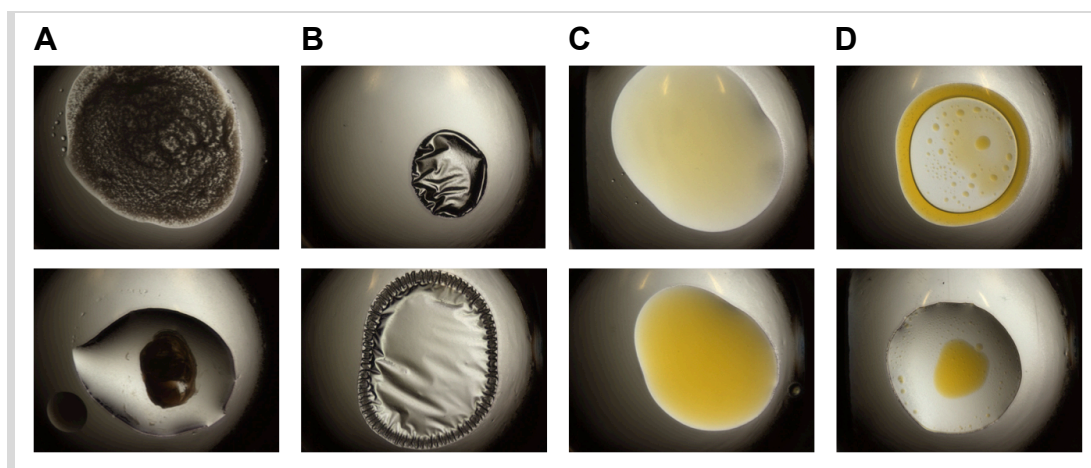


Figure 75: Exemples représentatifs des gouttes obtenues après un mois de criblage de conditions de cristallogenèse. A) Protéines précipitées. B) Formation de pellicule. C) Protéines solubles. D) Séparations de phases.

Dans l'attente de pouvoir accomplir un nouveau criblage de cristallogenèse, pour lequel le délai d'obtention de résultats positifs est difficile à estimer, nous avons réalisé une étude de cartographie de l'interaction NF5:GFP par mutations en alanine des résidus clés à la surface de la GFP. Des variants de GFP ont donc été générés par méthode *Quick-change* (Agilent-Technologies, 1991), validés par séquençage, exprimés en bactéries et purifiés par chromatographie d'affinité via leur étiquette N-terminale Strep-TagII. A l'exception du variant N164A qui n'a pas été exprimé correctement malgré plusieurs tentatives, un test ELISA a été effectué comme lors du criblage des tétramères de Nanofitines en utilisant la GFP non mutée ainsi que ses variants E95A, F99A, E101A, D102A, K162A et K166A. Les résultats obtenus (**Figure 76A**) ont mis en évidence une fixation spécifique de NF5 tétramérique à 20 μM en présence de GFP sauvage (WT), tandis qu'un signal de l'ordre du bruit de fond a été observé en absence de GFP (GFP-, correspondant à environ 4% du signal en présence de GFP). Une diminution à 4% et 10% du signal de référence a également été obtenue lors de l'utilisation des mutants F99A et K162A de la GFP, respectivement. Les autres mutations testées ont un impact moins important sur le signal ELISA, avec des écart-types plus importants et des signaux de 55% (K166A), 65% (E101A), 90% (E95A) et 110% (D102A). De façon intéressante, la transposition de ces résultats sur la structure prédite du complexe NF5:GFP (**Figure 76B**) indiquerait que les mutations affectant le plus la capacité de fixation de NF5 correspondent au résidu central de l'interface (F99) ainsi qu'à un des trois autres résidus

périphériques indiqués comme primordiaux par les scores de Rosetta (K162, mais pas significativement D102 ou E95).

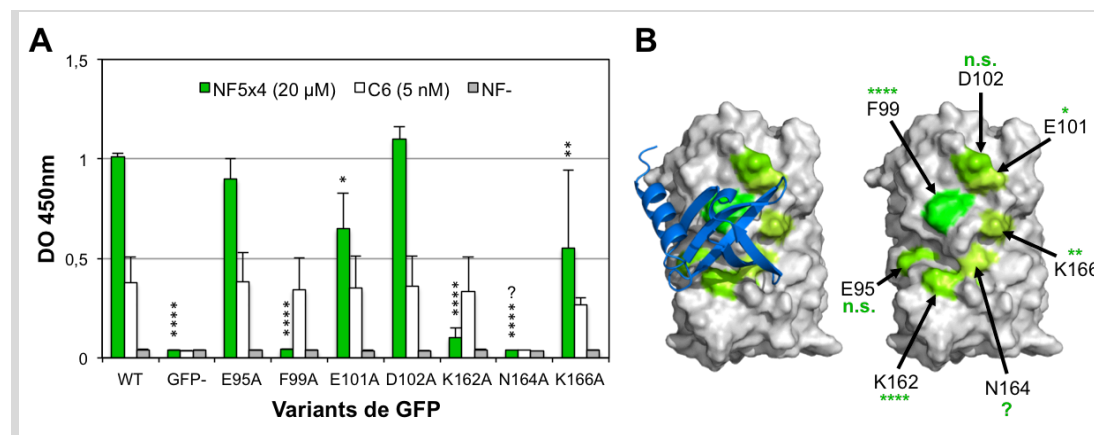


Figure 76: Cartographie de l'interaction NF5:GFP, prédite par mutations en alanine ou Rosetta. ****: $P < 0,0001$. **: $P < 0,01$. *: $P < 0,1$. n.s.: Non significatif. ?: Variant GFP mal exprimé ou non fonctionnel (absence de fluorescence). A) Résultats d'ELISA avec 20 µM de NF5 tétramérique (barres vertes), avec 5 nM de Nanotifine anti-GFP C6 (barres blanches), ou sans Nanotifine (barres grises) en présence des variants de StrepTagII-GFP immobilisés. WT: GFP non mutée. GFP-: Absence de GFP. B) Représentation de NF5 (cartoon bleu) et de la surface de la GFP (gris). Les résidus avec la plus grande contribution énergétique, prédite dans l'interaction par Rosetta, sont indiqués par une flèche et colorés en gradient de vert proportionnellement à leur importance calculée.

Il semblerait donc que l'interface que nous avons caractérisée en ELISA lors de la cartographie par mutations en alanine chevauche l'interface prédite. Sur quatre résidus primordiaux suggérés par les calculs de Rosetta, deux ont été confirmés par cette expérience, dont le résidu central de l'interface prédite à la surface de la GFP. Nous en avons conclu que l'épitope ciblé lors des prédictions a très probablement été effectivement fixé par NF5 *in vitro*. De ce fait, NF5 constituerait le premier exemple décrit de Nanotifine spécifique découverte intégralement par modélisation.

Ces résultats encourageants doivent tout de même être modulés sur deux aspects. D'une part, la validation rigoureuse de cette approche prédictive nécessite une confirmation de l'orientation du complexe à l'échelle atomique, demandant donc de cristalliser le complexe et d'en résoudre la structure tridimensionnelle. D'autre part, nous avons eu recours à la tétramérisation de NF5 et utilisé des concentrations de Nanotifine au delà du micromolaire. Ces conditions n'ont pas permis de calculer l'affinité de NF5 pour la GFP, mais nous pouvons raisonnablement en déduire que celle de NF5 sous forme monomérique est extrêmement faible. Dès lors, une augmentation

conséquence de cette affinité est nécessaire pour pouvoir offrir une application à cette Nanofitine sélectionnée rationnellement.

III.6. Optimisation de l'affinité de NF5 par modélisation

III.6.1. Stratégie d'exploration au voisinage de la pose NF5:GFP

L'approche décrite dans ce manuscrit a permis d'identifier la Nanofitine anti-GFP NF5 à partir de simulations d'ancrage moléculaire puis de design protéique. Ce variant représente par conséquent la meilleure solution obtenue par nos prédictions, bien que souffrant d'une affinité trop faible pour une détection efficace l'activité de sa forme monomérique.

Nous avons abordé précédemment que la découverte de NF5 fait suite à la convergence de l'ancrage moléculaire vers deux pôles à la surface de la GFP, puis à la convergence vers une famille de Nanofitines homologues de NF5 (NF1 à NF5) sur un de ces pôles. La sélection de ces séquences explorées par design *in silico* a d'ailleurs révélé qu'elles provenaient d'un nombre restreint de poses de départ, soulignant une corrélation importante entre l'orientation des squelettes peptidiques et l'optimisation des chaînes latérales qui y sont présentées. Nous avons aussi montré que l'épitope visé à la surface de la GFP est compatible avec des interactions de type protéine-protéine, à travers la cartographie de l'interaction NF5:GFP (de faible affinité) et la structure résolue d'une interaction Nanobody:GFP (de forte affinité). Nous avons alors décidé de tenter de découvrir des Nanofitines anti-GFP plus affines que NF5 tout en ciblant ce même épitope favorable à l'interaction. De plus, l'exploration des prédictions au voisinage de NF5 est aussi un moyen de vérifier la robustesse des algorithmes engagés en cas de résultats similaires.

La démarche adoptée peut être résumée comme une intensification de l'exploration spatiale autour de NF5, tel que représenté sur la **Figure 77**. Les étapes décrites précédemment dans ce chapitre constituent le premier tour de sélection *in silico*, identifiant l'orientation et la séquence de la pose de NF5. Cette pose a ensuite été perturbée localement par des cycles de relaxation (autorisant de légers mouvements de squelette peptidique, avec des RMS $\leq 0,75$ Å), d'ancrage moléculaire (offrant la possibilité de mieux amarrer les deux protéines relaxées dans une poche d'un rayon de 0,5 Å via des translations et rotations), puis de design de l'interface de la

Nanofitine à proprement parler. Compte tenu de la convergence rapide des résultats de design lors du tour 1, nous avons proposé d'augmenter le nombre de simulations pour les étapes de perturbations (relaxation et ancrage) afin de diversifier les poses dont l'interface a été modifiée, ce qui constitue les tours 2.1 et 2.2, puis le tour 3. Dans tous les cas, chaque étape d'un tour a été alimentée par les poses présentant les meilleurs scores à l'étape précédente.

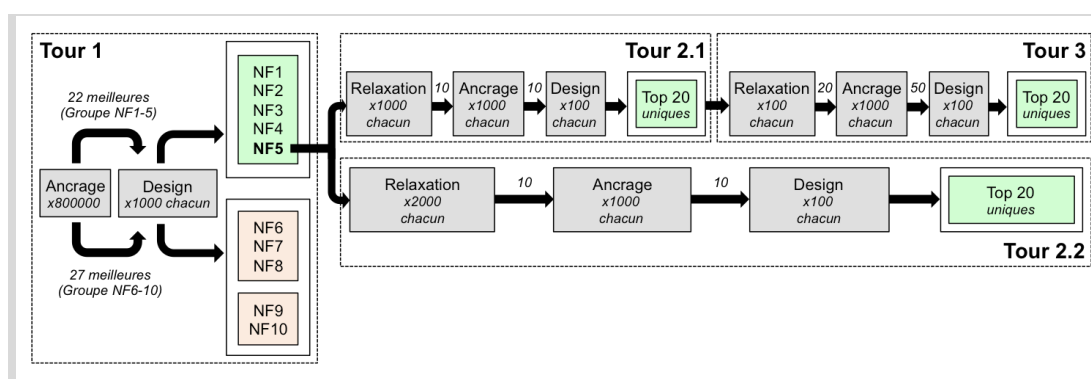


Figure 77: Représentation schématique des protocoles de design de novo de Nanofitines anti-GFP. Le tour 1 représente les conditions explorées ayant mené à l'identification des Nanofitines NF1 à NF5. Les tours 2.1, 2.2 et 3 ont été réalisés à partir de la pose de NF5 en complexe avec la GFP. Le nombre de résultats avec les meilleurs scores issus d'une étape et engagés dans l'étape suivante sont indiqués au niveau des flèches.

III.6.2. Convergence vers la famille de NF1 à NF5

Le tri par score d'interaction croissant des séquences uniques obtenues, retrouvées en fin de chaque branche (tours 2.2 et 3), a mis en évidence des séquences proches de celles des Nanofitines NF1 à NF5 retrouvées auparavant (**Figure 78**). En particulier, les meilleurs scores d'interaction ont été obtenus avec la séquence exacte de NF2 dans l'ensemble des conditions testées. Les séquences suivantes sont également des homologues de NF2, auxquelles s'ajoutent quelques mutations, notamment dans les meilleurs résultats du tour 3 (sous-représentés par rapport aux séquences avec moins de mutations). Les perturbations locales de NF5 auraient donc mené à la convergence vers NF2, qui avait déjà été identifiée et testée expérimentalement sous formes monomérique et tétramérique (via la conjugaison à la biotine et la complexation avec la streptavidine).

D'une part, ces observations signifient que NF2 serait la Nanofitine privilégiée par Rosetta dans l'environnement contraint qui lui a été proposé, et NF5 ne serait alors qu'un des variants de NF2. Or, NF2 a montré des signaux ELISA au moins deux fois plus importants en présence de GFP qu'en

absence de cible, ce qui pourrait être considéré comme le signe d'une Nanofitine anti-GFP de faible affinité. Des biais peuvent avoir été introduits dans ce criblage, notamment si des différences d'efficacité de multimérisation se sont produites, mais ces signaux ont été mesurés à des intensités plus faibles que celles avec NF5, signifiant que NF2 n'est sans doute pas significativement plus affine pour la GFP que ne l'est NF5 dans ces conditions.

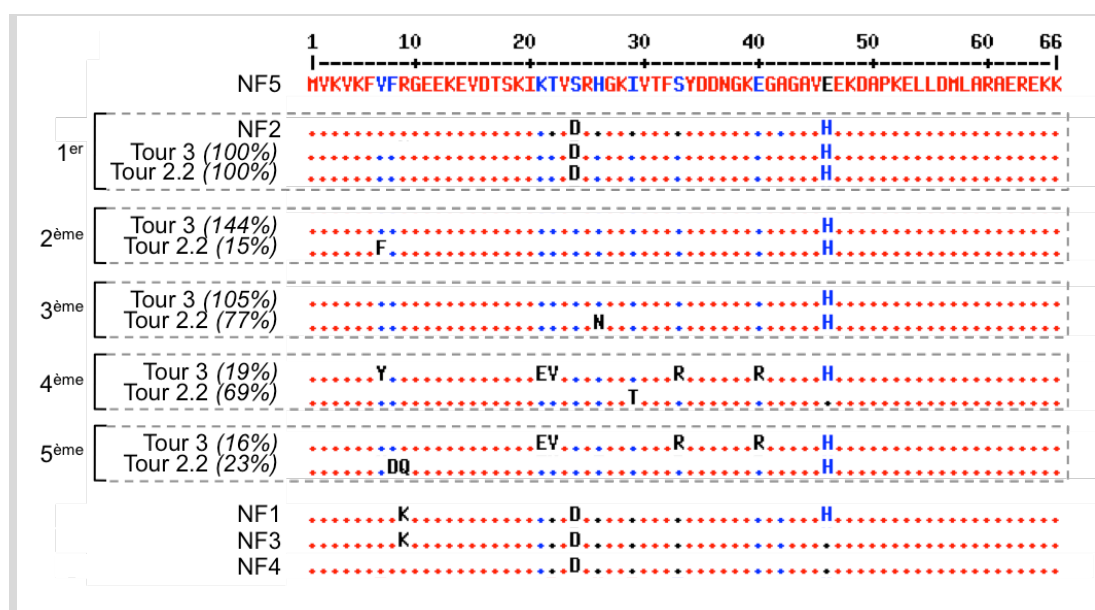
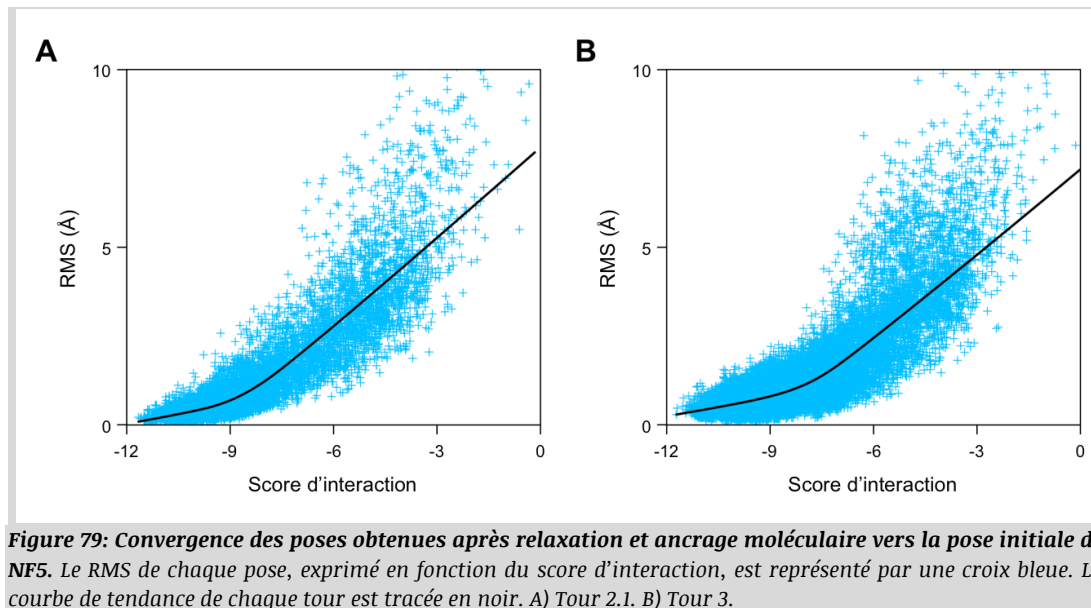


Figure 78: Alignement des 5 meilleures séquences obtenues lors des tours 2.2 et 3 de design, par rapport à NF5. Les pourcentages indiquent la fréquence de la séquence par rapport à celle de NF2 dans le même tour de sélection in silico. La séquence de NF2 est indiquée avec les meilleurs résultats de chaque tour (1^{er}) car ces séquences sont identiques. Les séquences de NF1, NF3 et NF4 sont indiquées pour rappel.

D'autre part, ces résultats sont indicateurs de la reproductibilité des prédictions fournies par Rosetta puisque l'approche que nous avons utilisée lors des tours 2 à 3 ne propose pas de solution effectivement différente de celles identifiées au préalable dans les meilleurs scores d'interactions du premier tour. L'analyse des RMS des poses générées pendant les tours 2 à 3, comparées à la pose de NF5 obtenue en fin de tour 1, a d'ailleurs révélé que les poses avec les scores d'interaction les plus négatifs sont celles déviant le moins de la pose de départ (**Figure 79**). Une raison de cette convergence pourrait donc provenir d'un environnement particulièrement avantageux pour les calculs de score de Rosetta. Ce phénomène pourrait d'ailleurs être entré en jeu dès l'étape d'ancrage moléculaire, puisque ses résultats ne présentaient pas de forte dépendance aux séquences exposées.



Finalement, la limite des outils que nous avons mis en place s'arrête à la découverte des Nanofitines homologues de NF2 et/ou NF5, présentant une affinité très modeste d'après la caractérisation effectuée sur NF5.

III.7. Conclusions et perspectives du design de novo de Nanofitines

III.7.1. Objectifs atteints

Au cours de ces travaux, nous avons tout d'abord identifié des épitopes compatibles avec la formation d'interactions protéine-protéine, grâce à des simulations d'ancrage moléculaire avec RosettaDock, dans une preuve de concept visant à fixer des Nanofitines à la GFP. Partant uniquement des structures de Sac7d et de la GFP, cristallisées indépendamment, nous avons été en mesure d'isoler deux pôles à la surface de la GFP pour y effectuer le design *de novo* de Nanofitines. La cohérence de ces prédictions de surfaces à cibler a par ailleurs été validée grâce à la résolution des structures de complexes Nanobody:GFP par Rothbauer et ses collaborateurs (Kirchhofer *et al.*, 2010). Nous avons également pu confirmer l'intérêt de la première banque de Nanofitines conçue, basée sur la mutation des 14 résidus impliqués dans la fixation de l'ADN par Sac7d, qui présenterait une orientation favorable à la fixation du tonneau beta de la GFP.

Une particularité de prédictions d'ancrage moléculaire présentées dans ce manuscrit réside dans la taille de l'espace à explorer. En effet, nous avons pu mettre en évidence la capacité de

RosettaDock à prédire des épitopes malgré une exploration de l'intégralité de la surface de la GFP. Pour des applications futures, nous anticipons une découverte plus rapide et exhaustive de poses favorables via une diminution du degré de liberté lors de l'ancrage moléculaire, notamment en définissant des régions restreintes à l'avance (sur la base des informations disponibles dans la littérature pour assurer le ciblage d'une interface protéique d'intérêt).

Ensuite, l'étude de design *de novo* décrite ici nous a mené à la découverte de la Nanofitine NF5, dont la tétramérisation a permis de mettre en évidence sa capacité de fixation spécifique à la GFP sur une interface chevauchant celle ciblée initialement. Cette interaction prédite différencie significativement notre approche des autres études portées à notre connaissance. La distinction réside principalement dans deux paramètres: l'utilisation des Nanofitines comme seule charpente protéique, et l'introduction des mutations prédites uniquement dans le site de fixation de la Nanofitine sans modification de la surface de la protéine cible. Toutefois, l'affinité de la Nanofitine anti-GFP générée est trop modeste pour pouvoir y attribuer une application en l'état, ce qui pourrait être compensé par des expériences de maturation d'affinité *in vitro*, quasiment systématiquement réalisées dans les études de design *de novo* documentées à ce jour (Whitehead *et al.*, 2013).

La convergence répétée vers des Nanofitines homologues de NF5 dans nos simulations de design, malgré une diversité autorisée dépassant $3,7 \times 10^{17}$ variants, témoigne de la robustesse et de la reproductibilité des résultats de RosettaDesign. Ceci semblerait aussi indiquer que la multiplication de poses différentes générées par ancrage moléculaire est une piste à explorer pour diversifier les résultats de design *de novo*.

Grâce à la redondance et à la convergence des résultats que nous avons obtenus lors de cet exemple de design rationnel de Nanofitines anti-GFP, cette étude a contribué à raccourcir le chemin à parcourir entre les prédictions de protéines d'affinité et l'obtention effective des propriétés attendues. A cette occasion, nous avons la possibilité de définir certaines limites de notre approche et d'y proposer des perspectives d'amélioration pour les applications à venir.

III.7.2. Extension des paramètres modélisés

L'approche que nous avons développée pendant cette étude semble ne pas pouvoir s'éloigner des séquences de NF2 ou NF5 pour fixer la région que nous avons visée à la surface de la GFP. Parmi les améliorations envisageables, il serait intéressant d'inclure des post-traitements indépendants des algorithmes de Rosetta pour s'affranchir d'éventuels biais méthodologiques (principalement dans les fonctions de scores employées). Nous pouvons notamment proposer d'utiliser des méthodes alternatives de calcul de scores d'interactions (autres que FoldX, qui n'a pas donné de résultats satisfaisants lors de nos essais car plus adapté pour des mutations ponctuelles; *Schymkowitz et al., 2005*). La recherche de cavités formées à l'interface a également été réalisée à l'aide de SurfaceRacer (*Tsodikov et al., 2002*) et 3V (*Voss et Gerstein, 2010*), sans permettre de discriminer efficacement les différentes poses générées (de manière attendue, le volume total des cavités augmente proportionnellement à la surface d'interaction).

Cependant, il pourrait être intéressant d'occuper ces espaces en ajoutant des molécules d'eau autour des interfaces et dans leurs cavités par dynamique moléculaire. Ces simulations pourraient compléter les modèles implicites de Rosetta et aider à identifier des liaisons hydrogène stabilisantes par exemple. Nous avons précédemment décrit l'interface NF5:GFP prédite comportant 3 liaisons hydrogène et étendue sur 894 et 837 Å², respectivement à la surface de la Nanofitine et de la GFP. A titre de comparaison, le Nanobody réducteur de fluorescence se fixant dans la même région (identifiant PDB: 3G9A) voit son interface réduite à environ 650 Å² mais bénéficie de l'établissement de 8 liaisons hydrogène. Sur l'autre pôle de la GFP identifié lors de l'ancrage moléculaire, le Nanobody amplificateur de fluorescence (identifiant PDB: 3K1K) possède également une affinité nanomolaire via l'établissement de 10 à 12 liaisons hydrogène sur 680 à 700 Å² de surface. L'augmentation de la surface de contact et la compatibilité géométrique observées dans notre étude pourraient nécessiter de renforcer les interfaces formées avec la formation de liaisons hydrogène supplémentaires.

Ces optimisations pourraient d'ailleurs être intégrées à Rosetta en modifiant les fonctions de scores utilisées, comme par exemple en utilisant une fonction alternative telle que *hpatch*, dédiée

à limiter l'exposition de résidus hydrophobes en surface (initialement développée pour augmenter la solubilité de protéines optimisées *de novo*; *Jacak et al., 2012*). Une étude menée par Procko et ses collaborateurs a démontré l'intérêt de ne pas cibler des régions hydrophobes d'une protéine lors du design d'inhibiteur d'enzyme, démontrant à l'occasion la difficulté accrue pour prédire des interfaces avec des résidus polaires (*Procko et al., 2013*). Lippow et ses collaborateurs ont également suggéré que la mise en place de modèles de potentiel électrostatique avec une précision accrue permettait une meilleure estimation des affinités prédites par rapport au seul calcul d'énergie libre lors de la formation de complexes (*Lippow et al., 2007*). Dans une comparaison de 5 succès et 158 échecs de design, Stranges et Kuhlman ont même suggéré que Rosetta pourrait ne pas équilibrer précisément les liaisons hydrogène et les énergies électrostatiques par rapport aux pénalités de désolvatation, et que le procédé de design pourrait ne pas explorer suffisamment de conditions pour identifier les conformations de chaînes latérales pouvant satisfaire pleinement le potentiel de liaisons hydrogène des interfaces (*Stranges et Kuhlman, 2013*). Intégrer l'évaluation de ces potentiels électrostatiques pourrait donc participer à modéliser plus précisément les phénomènes de solvation/désolvatation et d'établissement de liaisons hydrogène aux interfaces, en exploitant l'équation de Poisson-Boltzmann (*Fogolari et al., 2002*). Deux exemples de compilation de Rosetta intégrant des fonctions de scores personnalisées ont été décrits pour le calcul des constantes logarithmiques d'acidité de protéines (pKa; *Kilambi et Gray, 2012*) ou pour le développement d'inhibiteur du virus d'Epstein-Barr (*Procko et al., 2014*). Encore plus récemment, O'Meara et ses collaborateurs ont proposé une fonction de score alternative à la fonction standard *score12*, nommée *Talaris2014* ou *EleCHBv2*, qui semble pouvoir mieux refléter l'établissement de liaisons hydrogène et améliorer la prédiction de structure par Rosetta (*O'Meara et al., 2015*), suggérant que ce type d'optimisations est une voie à suivre pour le design *de novo*.

Pour finir, nous avons constaté avant design que l'impact des séquences des Nanofitines utilisées lors de l'ancrage moléculaire était négligeable par rapport à la conformation globale de la protéine. De ce fait, nous n'avons pas mis en place de "contre-ancrage" (ancrage moléculaire sur des régions différentes de celles optimisées, ou sur l'ensemble de la surface de la GFP) à partir des

poses optimisées lors du design. Un gain de sensibilité lors des simulations d'ancrage pourrait donc confirmer la spécificité locale de chaque optimisation proposée.

III.7.3. Approches alternatives

Au delà des améliorations techniques implémentées aux fonctions de calculs, nous pourrions proposer des stratégies alternatives pour rendre le design *de novo* de protéines, et en particulier de Nanofitines, plus accessible au regard des résultats que nous avons obtenus.

Tout d'abord, la découverte des pôles à cibler sur la GFP paraît fiable car elle semble en corrélation avec les complexes Nanobody:GFP documentés. Une approche consisterait donc à ne pas chercher à optimiser NF5 mais à continuer l'exploration de la région qu'elle cible en autorisant des perturbations locales plus importantes, et en particulier des rotations de plus grande amplitude autour de l'axe formé entre les barycentres de la Nanofitine et de la GFP. De cette façon, nous pouvons nous attendre à obtenir des solutions prédites qui divergent des homologues de NF2 ou NF5, aboutissant à des Nanofitines compétitrices de NF5. Cette démarche serait d'ailleurs plus représentative d'une application future du design *de novo* de Nanofitines sur une interface restreinte avec des simulations d'ancrage moléculaire plus exhaustives.

Dans le cas précis de la preuve de concept présentée dans ce manuscrit, nous pourrions bénéficier des résultats de la génération *in vitro* de Nanofitines anti-GFP. Par exemple, nous pourrions confirmer les épitopes qui peuvent être fixés par des Nanofitines via une étude de cartographie par mutations alanine à la surface de la GFP, nous assurant de cibler une région accessible. Ce type d'approche a déjà généré des succès avec des anticorps (Barderas *et al.*, 2008; T. Li *et al.*, 2014). Des mutations en alanine à la surface des différentes Nanofitines permettraient également d'obtenir des informations complémentaires quant à leur orientation. Un autre moyen de valider l'application de RosettaDesign aux Nanofitines pourrait se concentrer sur le re-design d'interfaces à partir de complexes de structures résolues par cristallographie aux rayons X, suivant les quelques exemples réussis qui se succèdent depuis une dizaine d'années (dont Clark *et al.*, 2006; Lippow *et al.*, 2007; Qiao *et al.*, 2012; Henager *et al.*, 2012; ou Luo *et al.*, 2014). Ces approches rendraient la génération de Nanofitines assistée par ordinateur plus accessible mais

constitueraient des déviations importantes par rapport à l'objectif initial de notre démarche, en proposant des optimisations de Nanofitines existantes plutôt que leur découverte directe.

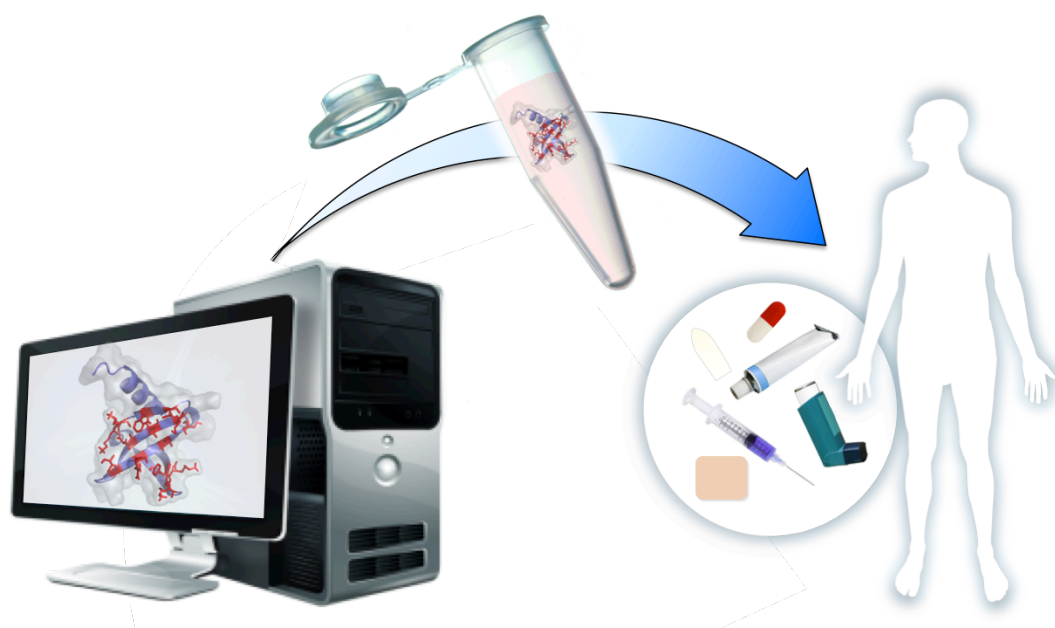
Indépendamment de la méthode d'obtention de l'orientation spatiale du complexe à optimiser, nous pouvons proposer de diminuer le clivage entre modélisation et expérimentation en augmentant le transit d'informations entre ces étapes complémentaires. D'une part, la méthode de sélection *in vitro* pourrait ainsi partager son taux de succès élevé de génération de protéines affines avec notre approche prédictive encore pauvre en réussites. D'autre part, la modélisation de complexes *de novo* ayant intégré les informations générées lors du processus de sélection *in vitro* pourrait donner accès à de meilleures protéines de liaison. Ces allers-retours entre paillasse et ordinateur pourraient se traduire par l'expression à petite échelle d'un nombre plus important de variants ponctuels à caractériser (produits par cultures bactériennes en plaques *deepwell* ou par traduction *in vitro*), ou en profitant de l'accessibilité récemment accrue des techniques de séquençage à très haut débit (Jain *et al.*, 2014) dont un exemple frappant a été appliqué au design et à l'optimisation de protéines anti-hémagglutinine (Whitehead *et al.*, 2012). Des échanges entre prédictions et expérimentations pourraient aussi être favorisés en ne limitant pas la validation des séquences identifiées par design aux meilleurs scores. Par exemple, ceci pourrait être réalisé via le criblage de sous-banques conçues pour refléter l'enrichissement observé *in silico*, tout en conservant des prédictions indépendantes afin de proposer une stratégie en entonnoir et identifier les conditions les plus favorables à la génération de Nanofitines affines et spécifiques.

De manière générale, nous anticipons également que les chances de succès de notre approche seraient rehaussées par la conception de banques de variants dirigée par les résultats de prédictions, tant au niveau des positions que des compositions en résidus à introduire. Dans le cas précis de notre étude, la sélection *in vitro* de Nanofitines anti-GFP via la randomisation de 11 résidus exposés sur le second feuillet beta de Sac7d a prouvé qu'elle pouvait générer des protéines d'affinité nanomolaire, tandis que l'ancrage moléculaire a confirmé une orientation globale des Nanofitines pour présenter ce même feuillet ainsi que la première boucle de Sac7d vers l'interface avec la protéine cible. De ce fait, les 14 positions de la banque utilisée lors des premières

génération de Nanofitines publiées semblent déjà optimisées et pourraient difficilement être étendues à des mutations additionnelles au voisinage de l'interface Nanofitine:GFP. En revanche, il serait envisageable de générer des banques de variants de Nanofitines possédant les résidus prédits en interaction avec la phénylalanine 99 de la GFP (localisée au centre de l'interface), voire également ceux au contact de la lysine 162 pour tenter d'imposer l'orientation du complexe. Cette stratégie de maturation d'affinité pourrait ainsi explorer une diversité réduite à travers une banque favorisant un épitope prédéfini. Pour valider cette méthode, il serait finalement possible d'effectuer un criblage du même type que celui qui serait réalisé lors d'une génération totalement *in vitro*, en évaluant l'impact de mutations en alanine à la surface ciblée de la GFP et/ou en réalisant une compétition avec une Nanofitine ou un Nanobody dont l'épitope est caractérisé.

Dans l'ensemble de ces propositions, la validation ultime des résultats obtenus consisterait à résoudre la structure des complexes par co-cristallisation aux rayons X. Nous nous accordons finalement à promouvoir un renforcement des échanges entre prédictions et expérimentations. D'une part, nous avons la conviction que cela donnerait plus aisément accès à des Nanofitines efficaces, aux propriétés permettant une application biotechnologique ou thérapeutique. D'autre part, cette intégration pluridisciplinaire pourrait alimenter les outils prédictifs, participer au développement des méthodes *in silico*, et réduire le nombre d'obstacles entre les modélisations appliquées aux interactions protéine-protéine et leur démonstration concrète.

Chapitre IV: Conclusions générales et perspectives



Chapitre IV: Conclusions générales et perspectives

IV.1. Ingénierie rationnelle des Nanofitines

Affilogic possède aujourd'hui un catalogue de protéines d'affinité dirigées contre plus de 40 cibles diverses (Prel, 2015), exploitant la robustesse de la charpente des Nanofitines, son fort potentiel de pénétration tissulaire et ses capacités de couplage ou fusion. Proposées comme alternatives aux anticorps lorsque ceux-ci montrent les limites inhérentes à leur structure complexe et/ou leur poids moléculaire élevé, les Nanofitines sont actuellement développées comme principe actif à application topique. Leur rôle thérapeutique est notamment assuré par leur capacité de neutralisation et de ciblage spécifique. Cette spécificité des Nanofitines est contrôlée lors de l'étape de sélection par *ribosome display*, réalisée au sein de la plateforme en place à Affilogic, qui donne efficacement accès à l'identification de variants de forte affinité. Aboutissant généralement à des K_D nanomolaires, le procédé n'affranchit cependant pas de l'effort de criblage consécutif à la sélection, qui vise à identifier les Nanofitines se liant à un épitope en faveur d'un effet thérapeutique. Nous avons donc souhaité augmenter le contrôle du ciblage précis des épitopes liés par des Nanofitines via une approche rationnelle de design par modélisation moléculaire. Supposant que prédéfinir le site de liaison d'une Nanofitine représente un moyen d'anticiper son mode d'action (spécialement dans le cadre de molécules bloquantes), ceci pourrait orienter la découverte de Nanofitines vers les protéines les plus efficaces.

Au delà d'un site de liaison muté, les Nanofitines tirent leurs propriétés de résistance et de simplicité de la charpente de la protéine archéobactérienne Sac7d, dont elles sont artificiellement dérivées. Du fait de cette origine non-humaine, Affilogic a souhaité investir dans un programme de recherche visant à explorer l'humanisation de ses molécules d'affinité. Le lancement de cette étude avait une vocation majoritairement préventive en matière d'immunogénicité puisque le risque de réaction immunitaire non désirée chez les patients était jugé faible compte tenu des propriétés physicochimiques des Nanofitines (avec des profils favorables en matière de taille, de solubilité et de stabilité). Finalement, cette démarche s'inscrit plus largement dans une volonté d'étendre le spectre des possibilités offertes par la génération de Nanofitines, en offrant

l'opportunité de les utiliser telles que découvertes *in vitro* ou comme support d'identification de sites de liaison (et donc comme précurseurs de protéines d'affinité humanisées à vocation thérapeutique).

Les approches de design *de novo* et d'humanisation par greffe de site de liaison ont été évaluées à travers le ciblage de la protéine de fluorescence verte (GFP) en mettant à profit une collaboration entre méthodes *in silico* et *in vitro*.

IV.2. Etat des avancées réalisées

IV.2.1. Transfert de site de liaison pour l'humanisation de Nanofitines

Dans le Chapitre II "Humanisation de Nanofitines par greffe de domaine", nous avons proposé une approche d'humanisation des Nanofitines pour anticiper d'éventuelles nécessités d'extension de demi-vie biologique et de réduction d'immunogénicité des Nanofitines. Le concept y a été envisagé de manière analogue aux greffes chirurgicales. Les résidus fonctionnels des Nanofitines (l'organe effecteur à transférer, ou greffon) ont été isolés depuis la charpente de Sac7d (le donneur) pour être greffés sur une charpente protéique humaine (l'hôte). Dans ce cadre, nous avons mis en place la greffe de 11 résidus sélectionnés sur une région rigide de la surface des Nanofitines vers des protéines humaines arborant un environnement jugé compatible. Cette compatibilité a été estimée lors d'un criblage de la PDB en quête d'une structure similaire au second feuillet beta de Sac7d, aboutissant à l'identification de 3 domaines protéiques humains.

Malheureusement, la greffe de deux feuillets anti-GFP différents sur ces charpentes humaines a entraîné l'insolubilité de deux d'entre elles et n'a pas permis de détecter d'activité de fixation à la GFP pour la dernière. N'ayant pas identifié de conditions expérimentales permettant de contourner ces phénomènes, nous avons proposé des pistes pour expliquer une partie de leurs déterminants via une analyse plus fine de l'environnement donneur et hôte des greffons (notamment en matière d'accessibilité). L'étude présentée dans ce manuscrit a néanmoins démontré un succès de transfert de charpente en réduisant les différences entre protéine donneuse et protéine hôte. Deux feuillets anti-GFP ont été greffés de Sac7d vers son homologue Sso7d, extrêmement proche sur le plan structural (avec un RMSD de seulement 0,40 Å sur

l'ensemble du squelette peptidique). Malgré la forte homologie entre les charpentes, le transfert a eu un impact sur l'affinité envers la GFP se traduisant par une augmentation de K_D d'un facteur 6. Ces résultats illustrent la finesse nécessaire au bon déroulement d'une greffe de charpente protéique sans maturation d'affinité additionnelle, tout comme l'avait souligné un autre exemple récent de transfert de Sac7d vers Sso7d (produisant une perte d'affinité d'un facteur 44; Béhar et al., 2014). Ces données précisent également l'importante perte d'affinité que nous pouvons anticiper en effectuant une greffe entre structures encore plus éloignées structurellement.

Par ailleurs, l'humanisation des Nanofitines s'est effectuée en concomitance avec d'autres travaux réalisés au sein d'Affilogic. Ces derniers ont abouti à la mise en place d'une technologie d'extension de demi-vie des Nanofitines et ont également démontré que la charpente de Sac7d induit une immunogénicité nulle ou négligeable. D'importance stratégique moindre, le pan d'optimisation des Nanofitines par humanisation a désormais été substitué par des routines de criblage des clones selon leur prédisposition à être présentés sur des complexes majeurs d'histocompatibilité de classe II (MHC-II) tout en guidant la désimmunisation du domaine variable des Nanofitines si nécessaire.

IV.2.2. Design rationnel assisté par modélisation

Composée d'étapes successives allant de l'ancrage moléculaire à la mutagenèse de la surface des Nanofitines, la stratégie de design *de novo* présentée dans le Chapitre III "Design *de novo* de Nanofitine spécifique" s'est principalement reposée sur l'utilisation de la suite de modélisation Rosetta. La simulation de complexes avec des scores de plus en plus favorables a permis de modéliser des interactions de type protéine-protéine entre la GFP et des Nanofitines. Les interfaces ainsi prédites, privilégiées sur deux pôles de la GFP, ont mis en avant une augmentation de l'aire de la surface d'interaction ainsi qu'une réorientation impliquant le second feuillet des Nanofitines lors du processus. L'exploration bioinformatique des espaces conformationnel et séquentiel étudiés a finalement aboutit à la caractérisation de 10 Nanofitines potentielles générées *de novo*. L'une d'elles s'est avérée avoir une activité de fixation spécifique à la GFP tout en souffrant d'une faible affinité (K_D de l'ordre de la dizaine de micromolaires sous

forme multiprésentée). L'épitope ciblé par la Nanofitine découverte *in silico* semble compatible avec les prédictions effectuées, dont les résidus cruciaux ont été confirmés par mutagénèse en alanine. Au delà de l'augmentation d'affinité nécessaire pour l'utilisation de la Nanofitine, l'orientation précise du complexe reste à élucider pour déterminer le degré de réussite de cette tentative de ciblage rationnel de Nanofitine par modélisation.

Par ailleurs, les étapes préliminaires de la méthode de design se sont avérées aptes à identifier des épitopes à fixer, coïncidant avec des interfaces connues entre la GFP et des protéines d'affinité. En plus de suggérer la validité des prédictions d'ancrage moléculaire, l'existence de ces ligands pourrait donner accès à une validation expérimentale complémentaire de la cartographie par mutation en alanine. La compétition entre Nanofitines prédites et ligands connus de la GFP pourrait aussi être évaluée avec les Nanofitines générées *in vitro* en amont de la stratégie d'humanisation des Nanofitines, à condition de déterminer plus précisément les épitopes qu'elles ciblent (par cristallographie aux rayons X ou à l'aide de variants alanine de la GFP).

Encourageants mais isolés, ces résultats pourraient être fortuits et nécessiteraient d'être reproduits avec d'autres séquences de Nanofitines dirigées contre la même cible et/ou via la fixation de cibles de nature différente. En plus de proposer une multiplication des cas d'étude, nous avons également identifié des verrous technologiques (essentiellement liés aux géométries des complexes avant design, ou aux fonctions de scores perfectibles quant à la modélisation de liaisons électrostatiques aux interfaces) qu'il serait bénéfique d'optimiser afin de réduire le risque d'échec encore élevé des approches computationnelles de design *de novo* (Stranges et Kuhlman, 2013).

IV.3. Défis et perspectives

A travers deux axes d'optimisation des Nanofitines, nous avons intégré des techniques de l'état de l'art de domaines complémentaires: la biologie moléculaire, la biochimie, la microbiologie et la modélisation moléculaire des protéines. Le dernier, en plein essor depuis les 30 dernières années, propose des méthodes attractives et théoriquement d'une extrême puissance sur les plans explicatif ou prédictif. Entre nos mains, ces outils ont généré des résultats encourageants mais

modestes, menant au transfert de sites de liaison entre deux homologues structuraux et à la découverte d'une Nanofitine de faible affinité par design *de novo*.

La modélisation moléculaire représente l'opportunité de guider rationnellement le développement de protéines aux propriétés désirées avec les biais inhérents à la traduction de phénomènes biophysiques en modèles théoriques (fonctions de scores imparfaites, algorithmes simplifiés non exhaustifs, etc.). Les quelques succès décrits à travers ce manuscrit et la littérature pointent la nécessité d'affiner les modèles existants. Cette vision structurale optimisée devra notamment mieux tenir compte des orientations favorables au trait souhaité (de l'échelle des complexes inter-moléculaires à celle de l'établissement de liaisons polarisées). Il serait également bénéfique de représenter plus fidèlement les phénomènes dynamiques modélisés, comme la flexibilité des protéines étudiées. En effet, l'accès à un degré de finesse suffisant pour prédire les mouvements des chaînes principales en complément de ceux des groupements fonctionnels pourrait se traduire dans une meilleure évaluation du repliement ou de l'activité de protéines de liaison (*Humphris et Kortemme, 2008*).

Par ailleurs, le chemin à parcourir ne réside pas uniquement dans les outils employés mais aussi dans le développement du savoir-faire qui leur est associé. Cette expertise est principalement enrichie au sein des groupes de chercheurs développant les outils bioinformatiques, comme l'illustrent parfois les résultats des évaluations CASP ou CAPRI comparant les serveurs de calculs en ligne et leur application avec intervention humaine. L'impact humain dans les méthodes automatisées d'exploration bioinformatique est d'ailleurs un axe d'amélioration possible de la modélisation moléculaire qui pourrait se traduire par l'intégration de l'intuition de l'expérimentateur dans le design protéique (*Lucas G Nivón et al., 2013*). Dans le cadre précis de l'étude présentée dans ce manuscrit, nous avons essentiellement adopté un positionnement en tant qu'utilisateurs d'applications innovantes de modélisation moléculaire et non en tant que développeurs de ces outils fondamentaux. Ainsi, notre rôle a consisté à mettre au point une démarche globale dans laquelle intégrer les outils disponibles en bénéficiant de leur développement fondamental (modifications des algorithmes, personnalisation des fonctions de

scores, etc.) et d'une volonté de partage par la communauté de développeurs (*Kaufmann et al., 2010; RosettaCommons.Org, 2015*).

Suite aux travaux menés à l'interface entre la modélisation et la validation expérimentale, nous encourageons fortement l'intensification des échanges entre les deux domaines. Nous anticipons que ces échanges permettront une modélisation de plus en plus précise, avec à la clé un possible accès à des prédictions avérées. Il serait alors intéressant d'étendre le champ des données générées pour favoriser l'accès à des exemples de succès dans une volonté d'apprentissage réciproque des techniques employées. Par exemple, cette ouverture pourrait se traduire par une convergence moins accentuée des modèles évalués expérimentalement, notamment via l'utilisation de banques éduquées de variants au lieu de mutants ponctuels. Une répartition plus équilibrée entre modélisation et expérimentation dans les premières phases de recherche pourrait accompagner graduellement nos travaux vers une compréhension plus fine des verrous technologiques et ainsi proposer des solutions efficaces pour les surmonter. Des approches graduelles pourraient par ailleurs réduire les difficultés que nous avons observées. Appliqué à la greffe de site de liaison, ce concept pourrait correspondre à une réduction du nombre de résidus à greffer (en caractérisant les protéines donneuses de manière approfondie) puis une incorporation séquentielle de mutations sur les charpentes hôtes. La diversification des modèles testés expérimentalement serait d'ailleurs plus accessible à l'heure actuelle compte tenu de l'augmentation de débit offerte par les progrès techniques et/ou de l'accès facilité à des technologies plus avancées au sein du laboratoire. A la fois source d'informations préalables et validation expérimentale ultime de nos travaux, l'obtention de structures cristallographiques de Nanofitines sous leur forme libre et/ou en complexe avec leur cible serait finalement un atout majeur. Il reste donc un effort conséquent à réaliser dans cette voie pour augmenter leur nombre et ainsi compenser le manque actuel d'une diversité des données structurales (principalement constituée de structures de Sac7d et Sso7d seules ou liée à un duplex d'ADN à l'heure actuelle). A terme, l'ensemble de ces perspectives pourrait participer à l'amélioration de notre compréhension et des outils à développer dans le cadre d'une recherche fondamentale.

D'autre part, les connaissances générées et consolidées par les travaux de thèse, par Affilogic ainsi que par des projets collaboratifs conduisent vers une caractérisation des Nanofitines en croissance constante. De ce fait, les défis surmontés et les nouveaux challenges ont été redéfinis depuis le lancement du projet de recherche décrit dans ce manuscrit. Alors que certains aspects comme l'humanisation des Nanofitines sont désormais délaissés, le chemin vers l'obtention de molécules thérapeutiques approuvées comporte toujours des axes d'optimisation essentiels. Parmi ceux-ci, figurent principalement le contrôle accru de la biodisponibilité, de l'absence de toxicité ou d'immunogénicité, ainsi que la maîtrise du mode d'action des Nanofitines. En règle générale, il est également évident que les applications des études prédictives devront dépasser le ciblage de la GFP. Nous espérons pouvoir focaliser ces outils avec succès sur des cibles d'intérêt thérapeutique pour compléter les améliorations intégrées *in vitro* et ainsi réduire le chemin à parcourir avant l'utilisation de Nanofitines comme médicaments.

Chapitre V: Matériels et Méthodes



Chapitre V: Matériels et Méthodes

V.1. Modélisations moléculaires

V.1.1. Recherche de feuillets compatibles dans la PDB

Pour identifier les protéines aptes à recevoir ces domaines anti-GFP, un criblage de la banque de données sur les protéines du *Research Collaboratory for Structural Bioinformatics (Protein Data Bank, ou PDB)* a été réalisé par Yves-Henri Sanejouand. En bref, les coordonnées spatiales du second feuillet beta de Sac7d ont été recherchées à partir des atomes de la chaîne principale (azote du groupe amine, carbone alpha et carbone du groupe carboxyle, respectivement désigné N, CA et C) des 11 résidus mutés (exposés en surface) ainsi que des 7 résidus enfouis (résidus V23, R25, V30, F32, G41, G43 et V45). La différence de géométrie a été estimée par calcul du RMSD entre ces atomes, avec les RMSD de plus faibles valeurs indiquant des géométries de feuillet semblables entre les charpentes donneuse et hôte. Les résultats considérés dans notre étude proviennent de la comparaison entre Sac7d et des structures de haute résolution ($R < 2$) de protéines humaines annotées monomériques et sans pont disulfure, avec une identité de séquence $\leq 90\%$.

Des indications supplémentaires sur les similarités d'environnement des feuillets identifiés lors du criblage de la PDB ont également été calculées. La différence d'accessibilité a été définie en calculant la valeur moyenne des surfaces accessibles par résidu, puis en déterminant le RMSD avec ces valeurs moyennes, devant indiquer les accessibilités voisines de celle du feuillet de Sac7d via les valeurs les plus faibles. La différence d'orientation des résidus a également été estimée par calcul de la similarité cosinus pour l'ensemble du feuillet beta ou uniquement le résidu le plus ré-orienté (cosinus le plus proche de 0 par rapport au feuillet recherché). Le calcul du cosinus de l'angle entre les directions des deux vecteurs permet d'identifier les cas de similarité maximale (correspondant à un angle de 0, soit un cosinus prenant la valeur de 1). Ainsi, la similarité cosinus (Cos) des résidus de feuillet correspond au produit scalaire de l'ensemble des vecteurs N→C (soit un vecteur à $18 \times 2 \times 3$ coordonnées), tandis que le pire cosinus (Cos min) obtenu pour un couple de vecteurs N→C a également été indiqué dans les résultats de criblage. Après identification des problèmes stériques rencontrés notamment avec la charpente 1LPJ, cette estimation de la similarité d'orientation a été affinée en calculant le cosinus des vecteurs orthogonaux au plan formé par les vecteurs N→CA et CA→C.

V.1.2. Analyses d'hydrophobie

Les différences d'hydrophobies prédites sous l'effet de l'introduction de feuillets anti-GFP sur les charpentes de Sac7d ou des structures 4F7H, 1LPJ, 1QNT et 1C8C ont été estimées à partir d'une échelle d'hydrophobie, initialement proposée par Yves-Henri Sanejouand et Georges Trinquier lors d'une étude de l'impact de mutations ponctuelles de bases nucléotidiques sur la conservation de

cette propriété (Trinquier et Sanejouand, 1998). En bref, les résidus ont été classés des plus hydrophobes (score maximum) vers les plus hydrophiles (score minimum) dans l'ordre suivant: WCMFILVGRS ATP EDKNQHY. La différence de score entre les charpentes après introduction des résidus fonctionnels des feuillets anti-GFP et les charpentes sauvages a ainsi été déterminée, définissant les feuillets induisant une augmentation d'hydrophobie (score > 0) ou une augmentation d'hydrophilie (score < 0) par rapport à la protéine sauvage.

V.1.3. Calculs de stabilité

V.1.3.1. Description de la méthode

Les prédictions de stabilité des différentes charpentes après introduction des résidus fonctionnels de feuillet anti-GFP issus de Nanofitines ont été réalisées par design *in silico* à partir des structures de 1AZP, 1C8C, 1LPJ, 1QNT et 4F7H préalablement relaxées. En imposant la séquence finale des variants simulés, l'algorithme employé est amené à résoudre un problème d'optimisation combinatoire pour identifier la conformation spatiale des groupements fonctionnels à exposer à la surface des squelettes peptidiques fixes. Cette exploration vise à atteindre un état conformationnel avec une énergie minimale, permettant ensuite de comparer les énergies totales avant et après introduction des mutations sur les protéines étudiées. Ces calculs ont été réalisés à l'aide de la suite logicielle Rosetta qui, à défaut de pouvoir garantir que les chaînes latérales proposées constituent les meilleures solutions, permet de réduire considérablement les temps de calcul nécessaires grâce à une approche stochastique de recuit simulé (Kirkpatrick, 1984; Černý, 1985) évitant le calcul préalable des énergies de toutes les paires de rotamères (Leaver-Fay et al., 2005, 2008). Les termes énergétiques et limitations de l'algorithme sont détaillés ci-après dans le cadre du design de la surface de Nanofitines, tel que décrit dans la section "V.1.4.4.2. Génération des poses de design".

V.1.3.2. Exécution des simulations

En bref, le protocole *fixbb* de la suite Rosetta 3.5 a été exécuté avec les paramètres suivants pour introduire les mutations et déterminer les scores d'énergie qui en découlent:

- *database*: base de données fournie par défaut
- *constant_seed* et *jran*: graine aléatoire fixée à 111111 pour des considérations de reproductibilité des calculs
- *nstruct*: 10 simulations ont été réalisées pour chaque variant étudié
- *s*: fichier de structure tridimensionnelle à muter pour estimer sa stabilité, préalablement relaxée avec l'application *relax* pour limiter les biais de score (tel que décrit dans la section "V.1.4.1. Préparation des poses")
- *resfile*: fichier listant les mutations à introduire sur la structure de départ
- *ex1* et *ex2*: augmentation de l'échantillonnage des rotamères

Après exécution à l'aide du script *queue.sh*, les scores d'énergie totale des différentes simulations ont été concaténés à l'aide du script *scores.sh* (Annexe 1.1 et Annexe 1.2, respectivement). Les fichiers *resfile* listant les mutations à introduire dans les différentes structures à évaluer ont été générés à la volée à partir du script *preparation.sh* (Annexe 1.3.1) et d'un fichier (*mutations-list-libX.txt*) comportant le nom des Nanofitines ainsi que les 11 résidus sélectionnés à la surface du feuillet anti-GFP (définie par %1% à %11% dans les modèles de fichiers *resfile*, Annexe 1.3.2 à Annexe 1.3.6).

V.1.4. Design de novo

Le design *de novo* de Nanofitines anti-GFP a été réalisé à partir de structures préalablement résolues pour générer des poses initiales relaxées pour chaque partenaire moléculaire. Les partenaires ont été orientés aléatoirement pour former un complexe lors de simulations d'ancrage moléculaire. Les meilleurs résultats ont alors été regroupés pour identifier les régions les plus favorables à une interaction de type protéine-protéine à la surface de la cible. Enfin, les poses sélectionnées à ce stade ont été utilisées pour introduire des mutations aléatoires sur la surface de la Nanofitine au niveau de l'interface entre les deux partenaires, aboutissant à l'identification de séquences de Nanofitines générées *in silico*.

V.1.4.1. Préparation des poses

V.1.4.1.1. Description de la méthode

Les poses des Nanofitines et de la GFP ont été préparées à partir des structures déposées dans la PDB sous les identifiants 1AZP et 3LVA, respectivement. Le motif AITHGMDELYK a été supprimé avant relaxation de la structure 3LVA, ce qui correspond à la suppression des atomes de l'extrémité C-terminale labile de la GFP. Les résidus cyclisés du fluorophore de la GFP ont également été omis, tout comme l'ensemble des atomes n'appartenant pas à la chaîne A de chaque structure.

Les poses ont ensuite été relaxées selon la méthode recommandée dans les démonstrations distribuées avec la suite logicielle Rosetta (en version 3.4 ou plus récente, d'après *Lucas Gregorio Nivón et al., 2013*), afin d'éliminer les rotamères de haute énergie qui pourraient introduire des biais erronés dans le classement des simulations d'ancrage moléculaire.

V.1.4.1.2. Exécution des simulations

En bref, le protocole *relax* de la suite Rosetta 3.4 a été exécuté avec les paramètres suivants:

- *database*: base de données fournie par défaut
- *relax:fast*: protocole simple de relaxation rapide sur 5 cycles (défaut)

- `relax:constrain_relax_to_start_coords` et `relax:coord_constrain_sidechains`: génère des contraintes en fonction de la structure cristallographique de départ pour la chaîne principale et les chaînes latérales, respectivement
- `relax:ramp_constraints`: désactivé, pour ne pas diminuer les contraintes au cours des cycles
- `s`: fichier de structure tridimensionnelle à relaxer
- `ex1` et `ex2`: augmentation de l'échantillonnage des rotamères
- `use_input_sc`: autorise l'utilisation des rotamères cristallographiques
- `flip_HNQ`: autorise les retournements du dernier angle χ des chaînes latérales des résidus histidine, asparagine et glutamine
- `no_optH`: désactivé, autorisant l'optimisation des liaisons hydrogène

Pour chaque structure, la relaxation a donc été exécutée via une commande de la forme:

```
[chemin vers les exécutable]/relax.linuxgccrelease -database [chemin vers la base de données] -relax:constrain_relax_to_start_coords -relax:coord_constrain_sidechains -relax:ramp_constraints false -s [chemin vers la structure PDB à préparer] -ex1 -ex2 -use_input_sc -flip_HNQ -no_optH false
```

Les structures relaxées ont été concaténées dans un fichier PDB unique, en prenant soin d'attribuer les chaînes A et B, respectivement, à la Nanofitine et la GFP. Les structures de variants de Sac7d exposant le feuillet anti-GFP de C6, C8, E2 (mélangés ou non) ou des résidus alanine ont été générées à partir de la pose Sac7d:GFP, de la façon décrite dans la section "V.1.3. Calculs de stabilité".

V.1.4.2. Ancrage moléculaire

V.1.4.2.1. Description de la méthode

Des simulations d'ancrage moléculaire ont été réalisées à partir des poses préparées pour Rosetta, générant 100 000 complexes par couple Nanofitine:GFP étudié. En résumé, chaque résultat de simulation obtenu par l'algorithme RosettaDock comporte plusieurs cycles de randomisation basés sur une méthode de Monte-Carlo en conservant un squelette peptidique fixe (*Chaudhury et al., 2011*). Deux types de cycles ont été appliqués après une perturbation initiale rendant totalement aléatoire l'orientation de la cible (rotations jusqu'à 180° par axe) et partiellement celle de la Nanofitine (rotations et translations jusqu'à 8° et 3 Å, respectivement).

Des cycles de perturbations (rotations et translations) gros-grain ont été effectués jusqu'à formation d'un complexe par la rencontre des deux partenaires moléculaires (dont les structures ont été simplifiées en remplaçant leurs chaînes latérales par leur centroïde). Pendant cette phase, une recherche de type Monte-Carlo a été effectuée en 500 étapes adaptant dynamiquement les rotations et translations appliquées pour atteindre un seuil d'acceptation de 25%.

Ensuite, des cycles de plus haute résolution ont été réalisés pour générer les complexes liés à partir du complexe présentant la plus faible énergie. Les représentations en centroïdes ont été remplacées par l'ensemble des atomes des chaînes latérales dans ces derniers cycles (sur la base des structures non complexées), afin d'affiner les simulations via l'exploration des rotamères en plus de celle des orientations globales des chaînes protéiques. Pendant cette phase, 50 étapes de recherche de type Monte-Carlo avec minimisation ont été réalisées. Durant ces cycles, les partenaires complexés avec squelette peptidique fixe ont été perturbés aléatoirement (selon une direction et une magnitude définies par une distribution gaussienne autour de 0,1 Å et 3°) avant minimisation des scores d'énergie lors de ces perturbations, puis les chaînes latérales ont été optimisées via l'exploration des rotamères suivie d'un test de critère de type Metropolis. Tous les 8 cycles, la compaction de l'ensemble des chaînes latérales a également été optimisée, suivie d'un test de critère de type Metropolis. En fin de cycles, une pose unique a finalement été retenue à partir du meilleur score d'énergie obtenu, représentant une proposition de complexe possible (ou pose d'ancrage moléculaire).

V.1.4.2.2. Exécution des simulations

En bref, le protocole *docking_protocol* de la suite Rosetta 3.4 a été exécuté avec les paramètres suivants:

- *database*: base de données fournie par défaut
- *s*: fichier de structure tridimensionnelle à perturber pour former un complexe
- *constant_seed* et *jran*: graine aléatoire fixée à une valeur différente (incrémentée par 1 à partir de 111111) entre chaque lot de simulations, permettant de diversifier les prédictions tout en assurant la reproductibilité des calculs
- *dock_pert*: perturbation initiale définie par une distribution gaussienne autour de 3Å et 8°
- *partners*: partenaires à ancrer, la valeur B_A définit la GFP (chaîne B) et la Nanofitine (chaîne A) comme premier et second partenaire, respectivement
- *spin*: autorise la rotation du second partenaire (la Nanofitine) autour de l'axe formé entre les barycentres des deux partenaires
- *randomize1*: randomisation de l'orientation du premier partenaire (la GFP)
- *ex1* et *ex2aro*: augmentation de l'échantillonnage des rotamères, tel que recommandé par défaut dans la documentation de Rosetta 3.4
- *nstruct*: 200 simulations ont été réalisées par lot de simulation (pour un total de 100 000 simulations distinctes par séquence de variant de Sac7d)

Pour chaque structure de départ, le protocole d'ancrage moléculaire a donc été exécuté via des commandes de la forme suivante:

```
[chemin vers les exécutables]/docking_protocol.linuxgccrelease
-database [chemin vers la base de données] -s [chemin vers la structure PDB de
départ] -constant_seed -jran [graine aléatoire] -dock_pert 3 8 -spin -nstruct 200
-out:file:scorefile [chemin vers le dossier de scores]/score.[N° de lot de
simulation]/.sc -mute core.io.database -no_filters -out:overwrite -randomize1 -ex1 -ex2aro
-out:path:pdb [chemin vers le dossier des fichiers PDB de sortie]
-out:suffix .[N° de lot de simulation] -docking:partners B_A >
[chemin vers le dossier de journalisation]/run.[N° de lot de simulation].log
```

Les poses ont ensuite été triées selon le score d'interaction (I_{sc}) des complexes générés, correspondant au score total du complexe duquel sont soustraits les scores de chaque partenaire isolé. D'après la documentation de l'application *docking_protocol*, des simulations de qualité satisfaisante sont généralement situées dans un intervalle de scores d'interactions compris entre -5 et -10. Nous avons donc considéré les meilleures poses d'ancrage moléculaire ($I_{sc} \leq -5$) pour effectuer le regroupement à la surface de la GFP.

V.1.4.3. Regroupement des poses

V.1.4.3.1. Analyse des Nanofitines et de leurs résidus impliqués dans l'interface

Les premières analyses de regroupement des poses générées par ancrage moléculaire ont été effectuées à l'aide du logiciel Calibur, particulièrement adapté pour analyser de nombreuses simulations simultanément (S. C. Li et Ng, 2010). Toutefois, ces analyses ont été effectuées à partir des structures globales des Nanofitines, possédant des surfaces d'interaction distinctes (avec des aires de contact situées sur différentes régions des Nanofitines) pouvant aboutir à un décalage entre le barycentre de la chaîne de la Nanofitine et le barycentre de l'interface formée. De plus, Calibur adapte par défaut le seuil de coupure appliqué au regroupement des structures, générant un paramètre plus difficile à normaliser entre les différentes conditions d'ancrage moléculaire. Nous avons donc réalisé les analyses suivantes sur un nombre restreint de poses tout en normalisant le seuil de coupure pour permettre la découverte de groupes de dimensions équivalentes. Les 200 meilleures poses d'ancrage moléculaire de chaque condition (présentant un $I_{sc} \leq -5$), ont été regroupées dans Gromacs (Daura et al., 1999) avec un seuil fixé à une valeur de 0,03 nm en tenant compte uniquement des résidus des Nanofitines au contact de la GFP.

V.1.4.3.2. Analyse des résidus de la GFP impliqués dans l'interface

En complément des représentations en groupes de barycentres des surfaces d'interactions, les résidus de la GFP engagés dans l'interface Nanofitine:GFP ont été identifiés en analysant les changements de coordonnées spatiales des carbones beta (atomes CB) avant et après formation du complexe par ancrage moléculaire. Les carbones beta de la GFP (chaîne B) ont été extraits de chaque fichier PDB généré par l'application *docking_protocol* de Rosetta, puis les mouvements observés ont été dénombrés et représentés sous forme de cartes de chaleur à la surface de la GFP.

V.1.4.3.3. Exécution des simulations

Les regroupements préliminaires réalisés avec Calibur ont été exécutés à partir de commandes de la forme suivante:

```
[chemin vers l'exécutable calibur] -c "AB" -t a [chemin vers la liste des fichiers PDB à regrouper] > [chemin vers le fichier de journalisation]
```

Les analyses de regroupement des surfaces d'interactions des Nanofitines ont été effectuées en générant un pseudo-atome aux coordonnées du barycentre de ces surfaces (en excluant les résidus 51 à 66 de l'hélice C-terminale de Sac7d) à l'aide du script python *interactions-clusters.py* (Annexe 1.4). Les paramètres ajustables dans ce script sont:

- *i*: fichier d'entrée, une liste de structures au format PDB à regrouper
- *m*: mode déterminant le type d'interface analysé et le format de sortie (par exemple, *m* = 2 analyse les barycentres de surface d'interaction uniquement et génère un état par pose dans un fichier PDB unique)
- *o*: fichier ou dossier de sortie, où stocker le(s) fichier(s) de sortie au format PDB
- *c*: rayon en Å de la sonde utilisée pour la détermination des surfaces
- *a*: sélecteur du partenaire Nanofitine, dont seuls les résidus d'interface sont conservés
- *b*: sélecteur du partenaire cible, dont l'ensemble des résidus est conservé
- *r*: sous-sélection du paramètre "a" à exclure de la définition de la surface d'interaction

Ce script a été exécuté dans l'environnement du logiciel de visualisation PyMOL (DeLano, 2002) via la commande suivante:

```
pymol -qc -r [chemin vers le répertoire de scripts python]/interactions-clusters.py -  
- -i='[chemin vers la liste des fichiers PDB à regrouper]' -o='[chemin vers le  
fichier PDB de sortie, ex:input.pdb]' -c=4.0 -a='chain A' -b='chain B' -r='resi 51-66'  
-m=2
```

Les fichiers générés (nommés *input.pdb* dans la commande précédente) ont été fournis comme entrée à l'application *g_cluster* de Gromacs, générant en sortie des groupes de poses (dans le fichier *output.pdb*) et la courbe de distribution des RMSD (dans le fichier *rmsd-dist.xvg*):

```
g_cluster -s input.pdb -f input.pdb -cl output.pdb -cutoff 0.03 -dist rmsd-dist.xvg
```

Les groupes identifiés ont ensuite été analysés manuellement et visualisés à l'aide du logiciel de visualisation PyMOL.

Les résidus de la GFP engagés dans l'interface Nanofitine:GFP ont été identifiés en analysant les changements de coordonnées spatiales des carbones beta de la GFP (chaîne B) ont été extraits de chaque fichier PDB généré par l'application *docking_protocol* de Rosetta, à l'aide de la commande suivante:

```
cat '[chemin vers fichier PDB à analyser]' | grep -e '^ATOM[\ ]\{1,\}[0-9]\{1,\}[\ ]  
[\ ]\{1,\}CB[\ ]\{1,\}[a-zA-Z]\{1,\}[\ ]\{1,\}B' > [chemin vers fichier de type PDB  
contenant uniquement les coordonnées des atomes CB]
```

Le dénombrement des modifications des coordonnées spatiales de chaque CB a été réalisé en exécutant la commande suivante dans l'environnement du logiciel de visualisation PyMOL, en renseignant les paramètres *in_name* (liste des fichiers PDB à analyser), *out_name* (fichier de sortie) et *ref_name* (fichier PDB de référence auquel comparer les positions de CB):

```
pymol -qrc ./pymol-CB-rms.py
```

La représentation des régions de la GFP les plus souvent modifiées a ensuite été réalisée à partir des commandes du fichier *.pymol* fourni par le script *pymol-CB-rms.py* (également à exécuter dans l'environnement du logiciel de visualisation PyMOL).

V.1.4.4. Design de la surface d'interaction des Nanofitines

V.1.4.4.1. Sélection des poses de départ

Avant de réaliser le design de l'interface par introduction de mutations sur la structure de Sac7d, nous avons identifié des poses chevauchant les groupes préférentiels identifiés lors des simulations d'ancrage moléculaire. Ces poses de départ (à fournir à l'application *fixbb* de Rosetta pour leur design *in silico*) ont été sélectionnées dès lors qu'au moins 5 carbones alpha (atomes CA) des résidus de la banque à 14 positions (Mouratou et al., 2012) ont été retrouvés à une distance inférieure à 10 Å du barycentre de l'épitope à cibler. De ce fait, le critère spatial de sélection employé a permis de tenir compte de la proximité des structures à muter (vis-à-vis de l'épitope ciblé) ainsi que de leur orientation vers le second feuillet beta de Sac7d.

V.1.4.4.2. Génération des poses de design

Le design *in silico* a été réalisé pour répondre au problème d'optimisation combinatoire permettant d'identifier les meilleurs groupements fonctionnels et leur conformation spatiale à exposer à la surface des squelettes peptidiques fixes pour stabiliser l'interaction. L'application *fixbb* de Rosetta a été exécutée pour résoudre ce problème NP-complet (dont il est difficile de trouver efficacement des solutions, bien qu'il soit possible de vérifier rapidement celles proposées; Pierce et Winfree, 2002), avec des temps de calcul considérablement réduits grâce à une approche stochastique de recuit simulé évitant le calcul préalable des énergies de toutes les paires de rotamères (Leaver-Fay et al., 2005, 2008). Les termes d'énergie utilisés lors du design à squelette peptidique fixe, calculés via la fonction *score12* sélectionnée par défaut dans l'application *fixbb*, sont listés dans le **Tableau 5**.

Du fait des approximations de fonctions et d'algorithmes engagées dans cette approche, les limitations suivantes sont à prendre en considération. Tout d'abord, ces prédictions ayant été réalisées sans exploration de l'espace conformationnel du squelette peptidique, les éventuels mouvements induits n'ont pas été pris en compte (bien que supposés négligeables dans le cas de l'interaction modèle entre des motifs de structures secondaires en feuillets beta). Ensuite, cette méthode propose une approche stochastique pour répondre à un problème NP-complet, ce qui implique que les résultats peuvent varier si le nombre de simulations réalisées est trop réduit

et/ou en utilisant des graines aléatoires différentes. Enfin, les fonctions de score utilisées ne pouvant être parfaites (Das et Baker, 2008), l'obtention de variants mal exprimés ou mal repliés reste possible (y compris lors de l'obtention de scores favorables).

Tableau 5: Termes d'énergie pris en compte lors du design par Rosetta avec la fonction score12.

Terme(s)	Description
fa_atr	Terme attractif de Lennard-Jones
fa_rep	Terme répulsif de Lennard-Jones
fa_sol	Energie de solvation de Lazaridis-Karplus
fa_intra_rep	Terme répulsif de Lennard-Jones intra-résidu
fa_pair	Terme statistique de paire, favorisant les ponts salins
fa_plane	Interactions pi-pi entre groupements aromatiques, par défaut = 0
fa_dun	Energie interne des rotamères dérivée des statistiques de Dunbrack
ref	Energie de référence de chaque acide aminé
hbond_lr_bb	Liaisons H entre chaînes principales, distantes dans la séquence primaire
hbond_sr_bb	Liaisons H entre chaînes principales, proches dans la séquence primaire
hbond_bb_sc	Energie de liaison hydrogène entre chaîne latérale et chaîne principale
hbond_sc	Energie de liaison hydrogène entre chaînes latérales
p_aa_pp	Probabilité de l'acide aminé aux angles phi/psi
dslf_ss_dst, dslf_cs_ang, dslf_ss_dih, dslf_ca_dih	Scores statistiques des ponts disulfures (distance S-S, angles et dièdres)
pro_close	Energie de fermeture du cycle de proline
rama	Préférences de Ramachandran
omega	Angle diédral omega dans la chaîne principale

Cette approche a été appliquée en autorisant la randomisation des résidus à la surface des Nanofitines aux mêmes positions que celles décrite lors de la première conception de banque (Mouratou et al., 2012), représentant 14 résidus pouvant être substitués par l'ensemble des acides aminés naturels, à l'exception des résidus prolines (risquant de modifier la conformation de la Nanofitine) et cystéines (augmentant le risque d'agrégation et pouvant établir des ponts disulfures masquant une partie du site d'interaction).

V.1.4.4.3. Exécution des simulations

Le dénombrement des CA de chaque pose retenue après ancrage moléculaire et l'identification des poses de départ suffisamment proches de l'épitope ciblé a été réalisé à l'aide du script python *interactions-resn-to-clusters.py* (Annexe 1.6). Les paramètres ajustables dans ce script sont:

- *i*: fichier d'entrée, une liste de structures au format PDB à analyser
- *c*: distance maximale en Å entre le pseudo-atome au barycentre de l'épitope ciblé et les CA
- *a*: sélecteur des résidus de la Nanofitine à conserver pour le dénombrement
- *e*: sélecteur du barycentre de l'épitope autour duquel effectuer le dénombrement
- *o*: fichier de sortie, listant les structures et leur nombre de CA à moins de "*c*" Å du pseudo-atome défini par "*e*"

Pour identifier les poses de départ du design de chaque épitope ciblé, ce script a été exécuté dans l'environnement du logiciel de visualisation PyMOL via la commande suivante:

```
pymol -qc -r [chemin vers le répertoire de scripts python]/interactions-resn-to-clusters.py -- -i='[chemin vers la liste des fichiers PDB à analyser]' -o='[chemin vers le fichier de sortie]' -c=10 -a='c. A and resi 7+8+9+21+22+24+26+29+31+33+40+42+44+46' -e='c. B and resi [N° du résidu au centre de l'épitope ciblé] and n. CA'
```

Les poses avec au moins 5 carbones alpha (atomes CA) des résidus de la banque à 14 positions retrouvés à une distance inférieure à 10 Å du barycentre de l'épitope à cibler ont été fournies comme entrée au protocole *fixbb* de la suite Rosetta 3.5, exécuté avec les paramètres suivants:

- *database*: base de données fournie par défaut
- *s*: fichier de structure tridimensionnelle à modifier pour optimiser le complexe
- *constant_seed* et *jran*: graine aléatoire fixée à 111111 pour des considérations de reproductibilité des calculs
- *ex1* et *ex2*: augmentation de l'échantillonnage des rotamères, tel que recommandé par défaut dans la documentation de Rosetta 3.5
- *nstruct*: 1000 simulations ont été réalisées par pose de départ (soit 27 000 et 22 000 pour les pôles 1 et 2, respectivement, ciblés à la surface de la GFP)
- *resfile*: fichier listant les mutations à introduire sur la structure de départ

Pour chaque structure de départ, le protocole d'ancrage moléculaire a donc été exécuté via des commandes de la forme suivante:

```
[chemin vers les exécutable]/fixbb.linuxgccrelease -database [chemin vers la base de données] -s [chemin vers la structure PDB de départ] -constant_seed -jran [graine aléatoire] -nstruct 1000 -resfile [chemin vers le fichier resfile] -ex1 -ex2 -out:overwrite -out:file:scorefile [chemin vers le dossier de scores]/scores.[nom de la structure PDB de départ].[N° de lot de simulation].sc -out:path:pdb [chemin vers le dossier des fichiers PDB de sortie] -out:suffix .[N° de lot de simulation] > '[chemin vers le dossier de journalisation]/run.[nom de la structure PDB de départ].[N° de lot de simulation].log'
```

La randomisation des résidus à la surface des Nanofitines a été autorisée aux mêmes positions que celles décrite lors de la première conception de banque (Mouratou et al., 2012), représentant 14 résidus pouvant être mutés en n'importe quel acide aminé naturel à l'exception des prolines et cystéines, comme indiqué dans le fichier *resfile* utilisé:


```

NATAA # allow only the natural amino acid, this default command applies to all residues that
are not given non-default commands
start
#Lib Y
7 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
8 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
9 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
21 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
22 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
24 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
26 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
29 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
31 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
33 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
40 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
42 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
44 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline
46 A NOTAA CP #Mutate the NF (chain A) with all AA except cysteine and proline

```

Les scores d'interaction (I_{sc}) des complexes optimisés par design ont ensuite été déterminés en recalculant les scores de chaque partenaire isolé afin d'en soustraire la somme au score total du complexe. Pour cela, les atomes des chaînes A et B ont été extraits dans des fichiers PDB séparés de celui du complexe généré par Rosetta, puis analysés par l'application *score* de Rosetta avec la fonction *score12* préalablement utilisée dans l'application *fixbb*.

Les profils de diversité de séquences de chaque condition de design (regroupée ou non par épitope ciblé, avant ou après tri des meilleurs scores) ont été visualisés sous forme de logos de séquences à l'aide du logiciel WebLogo (version 2.8.2; Crooks et al., 2004), en exécutant le script *weblogo-loop.sh* (Annexe 1.7). L'unique paramètre à renseigner (*i*) est une liste (en format tabulation) comprenant sur chaque ligne le chemin vers un fichier au format FASTA contenant les séquences, le titre du logo de séquence, et le nom du fichier à générer. La diversité des séquences avec les meilleurs scores de design a également été observée en alignant ces séquences à l'aide du serveur en ligne Multalin (version 2000/03/28; Corpet, 1988), en figurant les différences avec la séquence de Sac7d sauvage.

V.1.4.5. Aires de contact et cavités à l'interface

Les aires de surfaces d'interaction ont été estimées via Surface Racer (version 5.0; Tsodikov et al., 2002), à partir des structures isolées de la GFP et des Nanofitines (structure complète, ou feuillet beta uniquement) ainsi que les complexes formés lors de l'ancrage moléculaire ou du design. Les surfaces accessibles, déterminées avec le premier ensemble de rayons de van der Waals de Surface Racer et une sonde de 1,4 Å de rayon (modélisant une molécule d'eau), ont permis de définir les différentes surfaces d'interactions suivantes:

$$Interface\ totale = \frac{GFP + Nanofitine - Complexe}{2} \text{ ou } Interface\ via\ feuillet = \frac{GFP + Feuillet - Complexe}{2}$$

L'implication du feuillet dans la formation des interfaces a également été déterminée par le calcul du ratio $\frac{Interface\ totale}{Interface\ via\ feuillet}$

L'évaluation des complexes et partenaires isolés via Surface Racer a également identifié des cavités au sein des interfaces. Leur volume a ensuite été évalué à l'aide du logiciel 3V (Voss et Gerstein, 2010). Les calculs des aires et volumes ont été réalisés à l'aide du script *pdbsurf-auto.sh* (Annexe 1.8) acceptant des chemins de fichiers PDB comme paramètres, modifié à partir d'une version écrite par Yves-Henri Sanejouand.

V.1.4.6. Design complémentaire (après le tour initial)

V.1.4.6.1. Stratégie de perturbations locales dérivées de NF5

L'exploration spatiale autour de la pose NF5:GFP a été intensifiée suite au premier protocole de design protéique de Nanofitines (Tour 1, décrit dans la section "V.1.4.4. Design de la surface d'interaction des Nanofitines"), afin de découvrir des Nanofitines anti-GFP plus affines que NF5 tout en ciblant le même épitope favorable à une interaction protéine-protéine. La pose prédite de NF5 en complexe avec la GFP a été perturbée localement par des cycles de relaxation autorisant de légers mouvements de squelette peptidique, d'ancrage moléculaire dans une poche d'une rayon de 0,5 Å après translations et rotations, puis de design de l'interface de la Nanofitine. Le nombre de simulations a été augmenté lors des étapes de perturbations (relaxation et ancrage) afin de diversifier les poses de départ du design et ainsi favoriser l'obtention de séquences différentes de celle de NF5. Chaque étape de perturbation ou de design a été alimentée par les poses présentant les meilleurs scores à l'étape précédente de manière cyclique.

V.1.4.6.2. Exécution des scripts

Afin de pouvoir refaire rapidement le design décrit précédemment (désigné Tour 1) et des tours supplémentaires (nommés Tours 2.1, 2.2 et 3), le script *batch.sh* (Annexe 1.9) a été mis en place pour favoriser des cycles successifs et alternatifs entre ancrage moléculaire, relaxation et design. Les meilleurs résultats de chaque cycle ont ainsi été concaténés dans un fichier *.sum* qui liste les poses avec les meilleurs scores (après tri et dénombrement des séquences uniques lors de cycles de design). Ce fichier permet un accès synthétique aux résultats de cycles de simulations, tout en fournissant un fichier d'entrée pour d'éventuels cycles additionnels. La définition des cycles à simuler a été réalisée via le paramètre "rounds" qui concatène, entre virgules, les cycles successifs sous le format suivant:

```
[type de cycle (relax|docking|design)]:[nombre de poses totales à générer]:[nombre de poses à conserver en fin de cycle]
```

Par exemple, les résultats du Tour 2.1 ont été réalisés à partir de la pose de NF5 du Tour 1 à l'aide du script *batch.sh* avec l'option suivante:

```
rounds='relax:1000:10,docking:1000:10,design:100:20'
```

Les résultats du Tour 3 ont ensuite été obtenus à partir du fichier *.sum* généré lors du Tour 2.1, avec l'option suivante:

```
rounds='relax:100:20,docking:1000:50,design:100:20'
```

L'accès direct aux résultats du Tour 3 aurait également été possible en une instruction unique avec l'option suivante:

```
rounds='relax:1000:10,docking:1000:10,design:100:20,relax:100:20,docking:1000:50,design:100:20'
```

L'ensemble des paramètres ajustables du script sont les suivants:

- *amsterdam*: 1 (défaut) ou 0, pour définir si les scripts et applications sont exécutés sur le serveur de calcul *Amsterdam* ou en local
- *s*: pose(s) utilisée(s) lors du premier cycle (peut être une liste au format *.sum* ou une pose initiale au format PDB)
- *f*: répertoire de travail de Rosetta, dans lequel seront stockés les fichiers temporaires et les fichiers de sortie
- *o*: nom du fichier texte dans lequel seront listées les commandes à exécuter en bash pour permettre la parallélisation, situé dans le répertoire "*f*"
- *initial_offset*: 0 par défaut, à incrémenter avec le nombre de tâches déjà effectuées pour permettre de continuer une simulation fractionnée
- *simult_proc*: 1 par défaut, nombre de processeurs à solliciter simultanément
- *pert_deg*: 8 par défaut, angle de la perturbation initiale des cycles d'ancrage moléculaire (distribution gaussienne, en degrés)
- *pert_dist*: 3 par défaut, translation initiale des cycles d'ancrage moléculaire (distribution gaussienne, en Å)
- *spin*: 0 (par défaut) ou 1, rotation du second partenaire (chaîne A) autour de l'axe formé entre les barycentres des deux partenaires lors de l'ancrage moléculaire
- *rand1*: 0 (par défaut) ou 1, randomisation de l'orientation du premier partenaire (chaîne B) lors de l'ancrage moléculaire
- *rand2*: 0 (par défaut) ou 1, randomisation de l'orientation du second partenaire (chaîne A) lors de l'ancrage moléculaire
- *unitrans*: rayon en Å de la poche dans laquelle contraindre le partenaire perturbé lors de l'ancrage moléculaire (ignoré si 0)
- *rounds*: tours à effectuer, en indiquant le type de simulation, le nombre de poses à explorer et le nombre de poses à conserver (stockées dans le fichier au format *.sum*)
- *max*: 1000 par défaut, nombre maximum de poses générées par tâche (le nombre de tâches est adapté automatiquement au nombre de processeurs sollicités via "*simult_proc*", au nombre de poses demandé via "*rounds*", ainsi qu'à cette valeur)
- *rmsmin*: 0 par défaut, valeur minimale de RMS (des carbones alpha) acceptée pour les résultats conservés dans le fichier au format *.sum*

- *rmsmax*: 0.5 par défaut, valeur maximale de RMS (des carbones alpha) acceptée pour les résultats conservés dans le fichier au format *.sum*
- *mock*: 0 (par défaut) ou 1, pour simuler les tâches sans exécuter les simulations coûteuses en temps de calcul (pour recalculer les scores de poses déjà générées auparavant par exemple)

Les tours 2.1, 2.2 et 3 ont été exécutés avec les paramètres par défaut, à l'exception des valeurs *unitrans*, *pert_deg* et *pert_dist* définies à 0.5, 0.1 et 0.2, respectivement, pour générer des perturbations locales. Le traitement de ces résultats a également bénéficié d'améliorations du script de calcul des surfaces et volumes de cavités à l'interface (*pdbsurf-auto-v1.5.sh*; Annexe 1.10) par rapport au Tour 1.

V.2. Validations expérimentales

V.2.1. Liste des réactifs

V.2.1.1. Liste des oligonucléotides utilisés

Le **Tableau 6** liste les noms et séquences des amorces utilisées dans l'ensemble des constructions décrites dans la section "V.2. Validations expérimentales". Ces amorces y seront uniquement désignées par les noms figurant dans ce tableau.

Tableau 6: Oligonucléotides utilisés pour les différentes constructions de biologie moléculaire.

Nom	Séquence 5' vers 3'
1208_C8_24_rev	TAATAACTCTTTCGGGGCATCTTTCTCGGTACGCGGCCGGCCAGCTTGCCGTTGTCGTCGTA
1208_C8_Ala1_rev	CTTGCCGTTGTCGTCGTAGCCAAACGTCACCTCTTTGCCGAGACGCCCAACCGCCGATCTTACTAGTGCCACTTC
1208_C8_Ala10_rev	TAATAACTCTTTCGGGGCATCTTTCTCGGTACGCGGCCGGCCAGCTTGCCGTTGTCGTCGTA
1208_C8_Ala11_rev	TAATAACTCTTTCGGGGCATCTTTCTCGGTACGCGGCCGGCCAGCTTGCCGTTGTCGTCGTA
1208_C8_Ala3_rev	CTTGCCGTTGTCGTCGTAGCCAAACGTCACCTCTTTGCCGAGACGCCCAACCGCCAGATCTTACTAGTGCCACTTC
1208_C8_Ala4_rev	CTTGCCGTTGTCGTCGTAGCCAAACGTCACCTCTTTGCCGAGACGCCCAACCGCCAGATCTTACTAGTGCCACTTC
1208_C8_Ala5_rev	CTTGCCGTTGTCGTCGTAGCCAAACGTCACCGCTTTGCCGAGACGCCCAACCGCCAGATCTTACTAGTGCCACTTC
1208_C8_Ala6_rev	CTTGCCGTTGTCGTCGTAGCCAAACGTCACCTCTTTGCCGAGACGCCCAACCGCCAGATCTTACTAGTGCCACTTC
1208_C8_Ala7_rev	CTTGCCGTTGTCGTCGTAGCCAAACGTCACCTCTTTGCCGAGACGCCCAACCGCCAGATCTTACTAGTGCCACTTC
1208_C8_Ala8_rev	TAATAACTCTTTCGGGGCATCTTTCTCGGTACGCGGCCGGCCCTTGCCGTTGTCGTCGTA
AcGFP-D102A_For	CCATCTTCTTCGAGGCTGACGGCAACTACAAG
AcGFP-D102A_Rev	CTTGAGTTGCCGTCAGCCTCGAAGAAGATGG
AcGFP-E101A_For	GCACCATCTTCTTCGGGATGACGGCAAC
AcGFP-E101A_Rev	GTTGCCGTCATCCGGAAGAAGATGGTGC

AcGFP-E142A_For	GGGCAATAAGATGGCGTACAACACAACGCC
AcGFP-E142A_Rev	GGGCGTTGTAGTTGTACGCCATCTTATTGCC
AcGFP-E95A_For	GGCTACATCCAGGCGCGCACCATCTTC
AcGFP-E95A_Rev	GAAGATGGTGGCGCCTGGATGTAGCC
AcGFP-F221A_For	CGATCACATGATCTACGCCGCTTCGTGACCG
AcGFP-F221A_Rev	CGGTCACGAAGCCGGGTAGATCATGTGATCG
AcGFP-F99A_For	CAGGAGCGCACCATCGCCTTCGAGGATGACG
AcGFP-F99A_Rev	CGTCATCCTCGAAGGCGATGGTGGCTCCTG
AcGFP-K162A_For	GGCCAAGAATGGCATCGCGGTGAACCTCAAGATCC
AcGFP-K162A_Rev	GGATCTTGAAGTTCACCGCATGCCATTCTTGCC
AcGFP-K166A_For	CATCAAGGTGAACCTCGCGATCCGCCACAACATCG
AcGFP-K166A_Rev	CGATGTTGTGGCGGATCGCGAAGTTCACCTTGATG
AcGFP-N144A_For	GGGCAATAAGATGGAGTACGCCTACAACGCCAC
AcGFP-N144A_Rev	GTGGGCGTTGTAGGCGTACTCCATCTTATTGCC
AcGFP-N146A_For	GATGGAGTACAACCTACGCCGCCACAATGTGTAC
AcGFP-N146A_Rev	GTACACATTGTGGGCGCGTAGTGTACTCCATC
AcGFP-N164A_For	GAATGGCATCAAGGTGGCCTTCAAGATCCGCCAC
AcGFP-N164A_Rev	GTGGCGGATCTTGAAGGCCACCTTGATGCCATTC
AcGFP-S208A_For	CAGAGCGCCCTGGCCAAGGACCC
AcGFP-S208A_Rev	GGGTCTTGCCAGGCGCTCTG
AcGFP-Y143A_For	GGGCAATAAGATGGAGGCCAACAACGCCAC
AcGFP-Y143A_Rev	GTGGGCGTTGTAGTTGGCCTCCATCTTATTGCC
AmpRdel_KpnI_rev	CATCTAGGTACCCTGTGACACCAAGTTACTCA
AmpRdel_PstI_for	TAGTGACTGCAGTCCCCGAAAAGTGCCAC
CamR_KpnI_rev1	CATCTAGGTACCGGCACCAATAACTGCCT
CamR_PstI_for	TAGTGACTGCAGTTTTGGCGAAAATGAGACGT
Gdel3AA_Rev	GGAGTCCAAGCTCAGTCATTATTTCTCGGTTCCGC
Gdel3AA_Rev	GGAGTCCAAGCTCAGTCATTATTTCTCGGTTCCGC
GFP_B1_rev	TACAATGGATCCTTGAAAATAAAGATTTTCGTTACCCCTGTACAGCTCATCCATGCCGTGGGTGA
GFP_E1_for	CACACAGAATTCATTAAGAGGAGAAAATTAACATATGGTGGAGCAAGGGCGCCGA

V.2.1. Liste des réactifs

GFPlink10.3_Rev	TGCACCGCTGCCACCTGCACCGCCTTTGTACAGCTCATCCATGCC
GFPlink20.1_Rev	CAGAGGCACCCGAGCTCCCACCCGCACTGCCGCCCTTGTACAGCTCATCCATGCC
GFPlink5.1_Rev	GGAGCCTGCACTACCACCGCCTTTGTACAGCTCATCCATGCC
Glink10.3_For	GCAGGTGGCAGCGGTGCAGGTTCCGGTCAAGGTGAAATTC
Glink20.1_For	GCTCGGGTGCCTCTGGTGGCGGTGCAGGGTCAGGCGGATCAGTCAAGGTGAAATTC
Glink5.1_For	GGTAGTGCAGGCTCCGTC AAGGTGAAATTC
Glink5.1_For	GGTAGTGCAGGCTCCGTC AAGGTGAAATTC
Glink5.1_Rev	GGAGCCTGCACTACCTTTCTCGCGTTCCGC
Glink5.2_For	GGTGCAAGTGGTCCGTC AAGGTGAAATTC
Glink5.2_Rev	GGAACCACTGCACCTTTCTCGCGTTCCGC
Glink5.3_For	GGCTCCGGCGCATCCGTC AAGGTGAAATTC
Glink5.3_Rev	GGATGCGCCGAGCCTTTCTCGCGTTCCGC
Gpls01C_For	TAATGACTGAGCTTGGACTCC
Gpls01C_For	TAATGACTGAGCTTGGACTCC
Gpls01C_Rev	GGAGTCCAAGCTCAGTCATTAATTAAGCTTTTTCTCGCGTTCCGC
Gpls01C_Rev	GGAGTCCAAGCTCAGTCATTAATTAAGCTTTTTCTCGCGTTCCGC
Gpls01N_For	CTATGAGAGGATCGCATCACCATCACCATCACGGATCCGTC AAGGTG
Gpls01N_For	CTATGAGAGGATCGCATCACCATCACCATCACGGATCCGTC AAGGTG
Gpls01N_Rev	GTGATGCGATCCTCTCATAG
Gpls01N_Rev	GTGATGCGATCCTCTCATAG
Gpls02N_ForGS	CTATGGGATCCGTC AAGGTGAAATTC
Gpls02N_RevGS	GAATTTACCTTGACGGATCCCATAGTTAATTTCTCTCTTTAATGAATTCGTGTG
Gpls03C_For	TAATGACTGAGCCTGCAGTTGGAC
Gpls03C_Rev	GTCCAACCTGCAGGCTCAGTCATTA
HisGFPN_For	CTATGAGAGGATCGCATCACCATCACCATCACGGATCCATGGTGAGCAAGGGCGC
MBP_B1_rev	TACAATGGATCCTTGAAAATAAAGATTTTCGTTGTGTTATTGTTATTGTTGTTGTT
MBP_E1_for	CACACAGAATTCATTAAGAGGAGAAATTAACATGAAAAATCGAAGAAGGTAAACTGGTAATCT
NF2.2_VFK	GGATCCGTCAAGGTGAAATTCGTGTTCAAGGGCGAAGAAAAAGAAGTGGACACTAGTAAGATC
NF2.2_VFN	GGATCCGTCAAGGTGAAATTCGTGTTCAACGGCGAAGAAAAAGAAGTGGACACTAGTAAGATC
NF2.2_VFR	GGATCCGTCAAGGTGAAATTCGTGTTCCGTGGCGAAGAAAAAGAAGTGGACACTAGTAAGATC

NF2.2_VHD	GGATCCGTC AAGGTGAAATTCGTGCATGATGGCGAAGAAAAAGAGTGGACACTAGTAAGATC
NF2.3_FLDMMWWT	CTTGCCGTTGTCGTCGTACGTAAACCACACCCATTTGCCCATACGGTCAACTAAGAAGATCTTACTAGTGTCACCTTC
NF2.3_KAWWVIAS	CTTGCCGTTGTCGTCGTAGCTAAAGGCCACGATTTTGCCCAACGCCAAACGGCCTTGATCTTACTAGTGTCACCTTC
NF2.3_KGWHWHS	CTTGCCGTTGTCGTCGTAGCTAAAGTGCACCCATTTGCCGTGACGCCAAACGGCCTTGATCTTACTAGTGTCACCTTC
NF2.3_KGWWWAS	CTTGCCGTTGTCGTCGTAGCTAAAGGCCACGACTTTGCCCAACGCCAAACGGCCTTGATCTTACTAGTGTCACCTTC
NF2.3_KTDHITS	CTTGCCGTTGTCGTCGTAGCTAAAGTGCACGATTTTGCCGTGACGATCAACCGTCTTGATCTTACTAGTGTCACCTTC
NF2.3_KTSHITS	CTTGCCGTTGTCGTCGTAGCTAAAGTGCACGATTTTGCCGTGACGACTAACCGTCTTGATCTTACTAGTGTCACCTTC
NF2.4_EAAE	TAATAACTCTTTCGGGGCATCTTCTCCTCCACGGCGCTGCGCCTTCTTGCCGTTGTCGTCGTA
NF2.4_EAAH	TAATAACTCTTTCGGGGCATCTTCTCGTGCACGGCGCTGCGCCTTCTTGCCGTTGTCGTCGTA
NF2.4_VERH	TAATAACTCTTTCGGGGCATCTTCTCGTGCACACGGCCTTCGCCAACCTTGCCGTTGTCGTCGTA
NF2.4_VESH	TAATAACTCTTTCGGGGCATCTTCTCGTGCACGCTGCCTTCGCCAACCTTGCCGTTGTCGTCGTA
NF2.4_WRFH	TAATAACTCTTTCGGGGCATCTTCTCGTGCACGAAGCCACGGCCCCACTTGCCGTTGTCGTCGTA
NF2.4_WRYH	TAATAACTCTTTCGGGGCATCTTCTCGTGCACGTAGCCACGGCCCCACTTGCCGTTGTCGTCGTA
Nusa_B1_rev	TACAATGGATCCTTGGAAGTACAGGTTTTCTCGA
Nusa_E1_for	CACACAGAATTCATTAAGAGGAGAAATTAATATGAACAAAGAAATTTGGCTGTAGTTGA
Sclib2.1For	GGAGATATATCCATGAGAGGATCGCATCACCATCACCATCACCATCAGGATCCGTCAAG
Sclib2.2For	GGATCCGTC AAGGTGAAATTCNNSNNSNNSGGCGAAGAAAAAGAGTGGACACTAGTAAGATC
Sclib2.3Rev	CTTGCCGTTGTCGTCGTASNNAAASNNACSNNTTTGCCSNNACGSNNAACSNNSNNGATCTTACTAGTGTCACCTTC
Sclib2.4Rev	TAATAACTCTTTCGGGGCATCTTCTCSNNACSNNGCCSNNGCCSNNCTTGCCGTTGTCGTCGTA
Sclib2.5Rev	CCATATAAAGCTTTTCTCGCGTCCGCGACGCTAACATATCTAATAACTCTTTCGGGGC
SclibKYK2_For	CATCACGGATCCGTCAAGGTGAAATTCAAATATAAAGGCGAAGAAAAAGAGTGGACACTAGT
Trx_B1_rev	TACAATGGATCCTTGGAAGTACAAGTTCGTCGTCACGGCCAGGTTAGCGTCGAGGAA
Trx_E1_for	CACACAGAATTCATTAAGAGGAGAAATTAATATGAGCGATAAAATTAATCACCTGACTGA
CamR_PstI_for	TAGTGACTGCAGTTTTGGCGAAAATGAGACGT
CamR_KpnI_rev	CATCTAGGTACCGGCACCAATAACTGCCT
AmpRdel_PstI_for	TAGTGACTGCAGTCCCCGAAAAGTCCAC
AmpRdel_KpnI_rev	CATCTAGGTACCTGTGACACCAAGTTACTCA
PURE_For	GCGAATTAATACGACTCACTATAGGGAATTCATTAAGAGGAGAAATTAATCTATG
PURE_Rev	CTAGCTTGGATTCTCACCA
Qe30_For	CTTTCGTCTTCACTCGA

V.2.1.2. Autres réactifs

- 2xYT: pH 7,5; 16 g tryptone, 10 g yeast extract et 5 g NaCl dans 1 l d'eau ultrapure
- 6X Orange DNA Loading Dye, Fermentas, R0631
- Acrylamide/Bisacrylamide 30%, VWR, 1.00639.1000
- Agar, Dutscher, 777476
- Agarose, Sigma-Aldrich, A9539
- Albumin from bovine serum, Sigma-Aldrich, A7906
- Ampicilline, VWR, MOLEM77154623
- Anticorps anti-His6-HRP, Qiagen, 34450
- BamHI: Fast Digest BamHI, Fermentas, FD0054
- BugBuster 10X, VWR, 70921-4
- Casein sodium salt, Sigma-Aldrich, C8654
- Chloramphénicol, Sigma-Aldrich, C0378-25G
- Colonne de dessalage: PD-10, Dutscher, 17-0851-01
- Coomassie: PageBlue protein Staining Solution, Fisher Scientific, E993P4
- *d*-biotin, Sigma-Aldrich, B4501
- DMSO, Sigma-Aldrich, D8418
- DNase I, Qiagen, 79254
- dNTP mix 10 mM: dNTP mix 10 mM each, Fermentas, R0192
- *E. coli* BL21 pLysS: Bactéries compétentes Acella (BL21 DE3) pLysS, Coger, EDG-10773
- *E. coli* DH5 α LacI^q: Bactéries compétentes DH5 α LacIq, Life Technologies, 18288019
- *E. coli* SHuffle T7 Express lysY: Bactéries compétentes Shuffle T7 Express pLysY, New England Biolabs, C3030H
- Eau ultrapure: H2O mQ 18 M Ω , Veolia
- EZ link Sulfo-NHS-LC-LC-Biotin, Fisher scientific, 21338
- Gel Green 10000X dans H2O, VWR, 730-2961
- Glucose, VWR, 1.08342.1000
- Glycérol 100%, VWR, 24387.292
- H₂O₂ 30%, Sigma-Aldrich, H1009
- HABA-Avidine, Sigma, H2153-IVL
- HCl 12 M, Sigma-Aldrich, 84415-100ML
- HindIII: Fast Digest HindIII, Fermentas, FD0504
- IPTG, VWR, 437145X
- Kanamycine, VWR, A1493.0025
- Kit RNeasy MinElute Cleanup, Qiagen, 74204
- Kit TranscriptAid T7 High Yield Transcription, Thermo Fisher, K0441
- KpnI: Fast Digest KpnI, Fermentas, FD0524
- MassRuler™ Express LR Forward DNA Ladder, Fermentas, SM1263
- Mélange enzymatique TranscriptAid: TranscriptAid Enzyme Mix, issu du Kit TranscriptAid T7 High Yield Transcription, Thermo Fisher, K0441
- Méthionine, Sigma-Aldrich
- MgCl₂ 50 mM, Ozyme, F530L
- MgSO₄, Biolabs, M0254L
- NaCl, VWR, 1.06404.1000
- NTP: NTP mix 25 mM each, issu du Kit TranscriptAid T7 High Yield Transcription, Thermo Fisher, K0441
- O'Generuler Express DNA Ladder, Fermentas, SM1563
- PageRuler Plus Prestained Protein Ladder, Fermentas, SM1811

-
- Pastilles OPD, Sigma-Aldrich, P8287
 - PBS: Phosphate Buffered Saline, Sigma-Aldrich, P4417
 - PEG 8000, Sigma-Aldrich, 89510
 - Persulfate d'ammonium, VWR, 17-1311-01
 - Phosphatase alcaline: Enzyme Fast AP Thermosensitive Alkaline Phosphatase, Fermentas
 - Phosphate citrate buffer, Sigma-Aldrich, P4809
 - Phusion polymerase: Phusion DNA polymerase, Fermentas, F-530S
 - Plaque Nunc Maxisorp fond plat, VWR, 442404
 - Pré-mix de traduction sans méthionine: 1,25 ml M2 (ATP 40 mM, GTP 10 mM, AMPc 20 mM, acétylphosphate 600 mM); 1 ml ARNt (MRE600) 25 µg/µl; 1,25 ml Tris acétate 2 M à pH 7,5; 3,125 ml PEG 8000 40%; 1,75 ml acides aminés 10 mM (sans méthionine); 0,1 ml acide folique 10 mg/ml; 1,525 eau ultrapure; 5 ml glutamate de potassium 2 M; 2,82 ml acétate de magnésium 0,1 M; 9,09 ml anti-ssrA 200 µM
 - Protein desalting spin column, Fisher Scientific, 89849
 - Protein Free Blocking Buffer, ThermoScientific, 37570
 - PstI: FastDigest PstI, Fermentas, FD0614
 - PureYield™ Plasmid Miniprep System, Promega, A1223
 - RevertAid H Minus RT, Fermentas, EP0451
 - Ribolock ribonuclease inhibitor 40 U/µL, Fermentas, EO0381
 - Ribonucleic acid from baker's yeast, Sigma-Aldrich, 83853
 - RNeasy MinElute Cleanup Kit, Qiagen, 74204
 - RT Buffer, Fermentas, EP0451
 - Sall: Fast Digest Sall, Fermentas, FD0644
 - SDS, VWR, 442442F
 - Sera-Mag SpeedBeads NeutrAvidin Microparticles, Thermo Scientific, 78152104010150
 - Sera-Mag SpeedBeads Streptavidin Microparticles, Thermo Scientific, 661552104010150
 - Slide-A-Lyser Mini Dialysis, Fisher Scientific, 69580
 - Streptavidine, Sigma-Aldrich, S0677
 - T4 DNA ligase, Fermentas, EL0011
 - Tampon de chargement non réducteur 5X: Lane Marker Non Reducing Sample Buffer, Fisher Scientific, E6424R
 - Tampon FD 10X: Tampon Fast Digest 10X, Fermentas, 777472
 - Tampon Phusion 5X: Phusion Buffer HF (5X), Fermentas, F-518
 - Tampon T4 10X, Fermentas, EL0011
 - Tampon TranscriptAid 5X: 5X TranscriptAid Reaction Buffer, issu du Kit TranscriptAid T7 High Yield Transcription, Thermo Fisher, K0441
 - Tampon Tris-glycine SDS: Tris-Glycine-SDS PAGE Buffer 10X (composé de Tris 0,25 M, Glycine 1,92 M et SDS 1% (m/v), National Diagnostics) dilué au 1:10 en eau ultrapure
 - TBS: Tampon 20 mM Tris-HCl, 150 mM NaCl, pH 7.4
 - TBS-T-BSA: TBS 1X additionné de Tween 20 0,002% et de BSA 0,01%
 - TEMED, VWR, 1.10732.1000
 - Tris-HCl, Calbiochem, 648317
 - Tris, Sigma-Aldrich, T6066
 - Trizma Base, Sigma-Aldrich, T1503
 - Tryptone, Dutscher, 777472
 - TSS: 2,56 g PEG 8000; 0,25 g MgSO₄; 1 ml DMSO; Qsp 20 ml milieu 2xYT
 - Tween 20, Sigma-Aldrich, P9416
 - Wizard SV Gel and PCR Clean-up System, Promega, A9281
 - Yeast Extract, Dutscher, 777474
-

V.2.2. Sélection *in vitro* de Nanofitines anti-GFP

V.2.2.1. Biotinylation des antigènes

La StrepTagII-GFP biotinylée a été utilisée pour la sélection et l'identification des clones de Nanofitines. La biotinylation a été réalisée par incubation d'une solution de cible protéique à 110 μM en présence de 5 excès molaires de sulfosuccinimidyl-6-(biotinamido) hexanoate (Sulfo-NHS-LC-LC-Biotin, Pierce) en PBS (Sigma-Aldrich) dans la glace pendant 1 h. Le tampon de la protéine biotinylée a ensuite été échangé à l'aide de colonnes de dessalage (Protein desalting spin columns, Pierce) équilibrées en tampon 20 mM Tris-HCl, 150 mM NaCl, pH 7.4 (TBS). Le degré de biotinylation a été déterminé à environ 2 molécules de biotine par molécule de protéine, à l'aide d'un test 4-Hydroxyazobenzene-2-carboxylic acid (HABA/Avidine, Sigma-Aldrich).

V.2.2.2. Tours de sélection par *ribosome display* et isolement des clones

La banque combinatoire de Nanofitines a été préparée comme décrit précédemment (Mouratou *et al.*, 2007, 2012), en randomisant les positions 21, 22, 24, 26, 29, 31, 33, 40, 42, 44 et 46 de Sac7d (exposées sur son second feuillet beta). En résumé, la banque a été assemblée par deux réactions de PCR chevauchantes successives à partir d'oligonucléotides dégénérés encodant des triplets NNS (N = A, C, T ou G et S = C ou G). Enfin, une étape finale de PCR a ajouté les régions nécessaires au *ribosome display* aux extrémités 5' et 3' (Hanes *et Plückthun*, 1997). Les banques amplifiées par PCR ont été transcrites et la sélection a été réalisée à 4°C, comme décrit par Mouratou *et al.* (Mouratou *et al.*, 2007, 2012). Bien que l'ajout de l'oligonucléotide anti-ssrA au mix de traduction a été décrit comme augmentant la stabilité des complexes ternaires ARNm-ribosome-protéine (Schaffitzel *et al.*, 1999), aucun effet bénéfique n'a été observé dans nos mains (lors de sélections précédentes) et la sélection a été effectuée en son absence. Six tours de sélection ont été réalisés pour isoler des interacteurs de haute affinité. La pression de sélection a été ajustée en augmentant graduellement le temps des étapes de lavages (8 lavages de 30 secondes, 3 minutes et 15 minutes, respectivement pour les tours 1, 2 et 3, puis 4 lavages de 15 minutes suivis de 4 lavages de 30 minutes pour les tours 4 à 6), ainsi qu'en diminuant la quantité de protéine cible utilisée pour l'étape de sélection lors des deux derniers tours (15 pmol pour les tours 1 à 4, puis 3,75 et 0,93 pmol, respectivement pour les tours 5 et 6). Le matériel ADN amplifié après le sixième tour a été cloné entre les sites de restriction BamHI et HindIII du plasmide pAFG01, dérivé du plasmide commercial pQE-30 (Qiagen). Le mélange réactionnel de ligation a ensuite été transformé en souche *E. coli* DH5 α LacI^q (Invitrogen). Les clones sélectionnés sur gélose de milieu 2xYT additionné de 100 $\mu\text{g}/\text{ml}$ d'ampicilline et de 25 $\mu\text{g}/\text{ml}$ de kanamycine ont été inoculés dans une plaque deep-well contenant 0,75 ml de milieu 2xYT additionné de 100 $\mu\text{g}/\text{ml}$ d'ampicilline, de 25 $\mu\text{g}/\text{ml}$ de kanamycine et de 1% de glucose par puits. Après culture à 37°C sous agitation à 600 rpm pendant la nuit, 0,2 ml de chaque culture a été inoculé dans une autre plaque deep-well contenant 1,25 ml de milieu 2xYT additionné de 100 $\mu\text{g}/\text{ml}$ d'ampicilline, de 25 $\mu\text{g}/\text{ml}$ de

kanamycine et de 0,1% de glucose par puits. La plaque a été incubée à 37°C pendant 3 heures sous agitation à 600 rpm. L'expression des clones de Nanofitines a été induite par ajout de 50 µl d'Isopropyl β-D-1-thiogalactopyranoside (IPTG) à une concentration finale de 0,5 mM et incubation à 30°C pendant 4 heures sous agitation à 600 rpm. Les cellules ont été collectées par centrifugation pendant 20 minutes à 2000g, et les surnageants ont été éliminés. Les protéines ont été extraites par ajout de 100 µl de réactif d'extraction protéique BugBuster IX (Novagen) par puits, suivi d'une agitation pendant 1 heure à température ambiante, avant ajout de 350 µl de TBS. Les débris cellulaires ont été culottés par centrifugation pendant 20 minutes à 2000g et les surnageants ont été engagés dans le processus de criblage.

V.2.2.3. Criblage des Nanofitines anti-GFP par ELISA

La streptavidine (100 µl à 66 nM, Sigma-Aldrich) en TBS a été immobilisée dans les puits d'une plaque adhérente Maxisorp (Nunc) par incubation à 4°C pendant la nuit. Toutes les étapes suivantes ont été réalisées à température ambiante, avec une agitation à 600 rpm pour les étapes d'incubation. Les puits ont été lavés 3 fois avec 300 µl de TBS, puis bloqués avec 300 µl d'albumine bovine sérique (BSA, Sigma-Aldrich) en TBS pendant 1 heure. Le contenu des puits a été éliminé par retournement rapide et l'immobilisation de la StrepTagII-GFP biotinylée (100 µl à 40 nM) a été réalisée par incubation pendant 1 heure en TBS additionné de BSA 0,5%. Avant chacune des étapes d'incubation suivantes, les puits ont été lavés 3 fois avec 300 µl de TBS contenant 0,1% de Tween 20. Les extraits bruts bactériens (100 µl, dilués à 1:40 en TBS contenant 0,1% de Tween 20) ont été mis dans des puits avec ou sans antigène immobilisé pendant 1 heure. La révélation a été effectuée par addition de 100 µl d'anticorps anti-RGS His conjugué à la HRP (Qiagen) dilué au 1:4000 en TBS contenant 0,1% de Tween 20 pendant 1 heure, suivie de l'ajout de 100 µl d'une solution de substrat *o*-Phenylenediamine dihydrochloride (OPD, Sigma-Aldrich) à 1 mg/ml en tampon de révélation (0,05 M acide citrique, 0,05% peroxyde d'hydrogène). L'absorbance à 450 nm a été mesurée à l'aide d'un lecteur de plaque ELISA Varioskan (Thermo Scientific).

V.2.2.4. Criblage complémentaire par interférométrie de couche biologique

Les profils cinétiques de fixation à la GFP et les taux de productions en microcultures ont été estimés par interférométrie sur un système Octet RED96 (ForteBio). Des biocapteurs Ni-NTA ont été équilibrés pendant 3 minutes en lysat sans Nanofitine dilué au 1:2 dans du TBS IX additionné de Tween 20 0,002% et de BSA 0,01% (TBS-T-BSA), puis les lysats bactériens à cribler dilués au 1:2 dans du TBS-T-BSA ont été chargés sur les capteurs pendant 5 minutes. Les cinétiques d'association et dissociation ont ensuite été évaluées en exposant les capteurs chargés des Nanofitines, fusionnées à une étiquette poly-histidine, à une solution de StrepTagII-GFP à 50 nM en TBS pendant 10 minutes puis à une solution de TBS pendant 10 minutes. Les biocapteurs ont été régénérés par incubation pendant 1 minute en NiSO₄ à 100 mM après 3 cycles de 5 secondes en glycine 10 mM à pH 2 et en TBS IX. Toutes les étapes ont été effectuées à 30°C avec une agitation

continue à 1000 rpm. Les sensorgrammes ont été analysés à l'aide du logiciel Octet Data Analysis version 7.1 (ForteBio), en utilisant un modèle d'ajustement 1:1 pour les étapes d'association et dissociation. L'étape de chargement des biocapteurs a également été analysée séparément en comparaison avec une gamme étalon réalisée dans les mêmes conditions (avec la Nanofitine H4 anti-lysozyme (Cinier *et al.*, 2009) diluée en lysat et TBS-T-BSA à une concentration finale de 200, 100, 50, 25, 12,5, 6,25, 3,125 et 0 µg/ml), pour estimer le taux de production de chaque variant de Nanofitine. Les taux d'expression estimés et les k_{off} apparents ont été pris en compte pour sélectionner les candidats à exprimer en erlenmeyer et à purifier.

V.2.3. Constructions de plasmides

V.2.3.1. Isolement, amplification et purification de plasmides

V.2.3.1.1. Transformation bactérienne et étalement

L'isolement des plasmides (préalablement purifiés, ou issus de mélange réactionnel de ligation ou d'assemblage Gibson) a été assuré par transformation de bactéries compétentes puis étalement sur géloses. Les bactéries transformées ont été obtenues à partir d'aliqots de bactéries déjà compétentes (stockés à -80°C) ou rendues compétentes extemporanément. En bref, les bactéries à rendre compétentes ont été inoculées dans 10 ml de milieu 2xYT additionné de 25 µg/ml de kanamycine (antibiotique pour lequel la souche exprime un gène de résistance), puis incubées à 37°C sous agitation jusqu'à atteindre une densité optique à 600 nm comprise entre 0,4 et 0,5. Les bactéries en suspension ont alors été aliqotées par 1 ml puis centrifugées pendant 5 minutes à 4500g. Les surnageants ont été éliminés et les culots bactériens stockés dans la glace jusqu'à la chimioporation réalisée par resuspension avec 200 µl de tampon TSS stérile.

Les plasmides (~100 ng) ont été ajoutés aux 200 µl de bactéries resuspendues en TSS ou à 50 µl de bactéries compétentes stockées à -80°C, puis la transformation a été réalisée par choc thermique par incubation pendant 30 minutes dans la glace, 50 secondes à 42°C puis 2 minutes dans la glace. Ensuite, 800 µl de milieu 2xYT sans antibiotiques ont été ajoutés aux bactéries transformées avant leur incubation à 37°C sous agitation pendant 1 heure. Les bactéries transformées ont alors été centrifugées pendant 5 minutes à 4500g. Les surnageants ont été éliminés par retournement et les culots bactériens resuspendus dans le reste de milieu, avant d'avoir été étalés sur des géloses de milieu 2xYT additionné de 100 µg/ml d'ampicilline et 25 µg/ml de kanamycine puis incubés pendant la nuit à 37°C.

V.2.3.1.2. Miniprep et caractérisation des plasmides

L'amplification des plasmides a été réalisée par inoculation de colonies bactériennes isolées dans 10 ml de milieu 2xYT additionné de 25 µg/ml de kanamycine (antibiotique pour lequel la souche exprime un gène de résistance) et de 100 µg/ml d'ampicilline (antibiotique pour lequel le plasmide procure la résistance), puis incubation à 37°C sous agitation pendant la nuit. Un aliqot

de chaque préculture a été préparé par addition de 370 µl de glycérol à 75% dans 1 ml de milieu de préculture saturant, puis a été conservé à -80°C. Les bactéries du volume restant de préculture ont été collectées par centrifugation pendant 15 minutes à 3220g. Les surnageants ont été éliminés et les bactéries resuspendues dans 600 µl d'eau ultrapure pour en extraire l'ADN plasmidique à partir du kit de miniprep "PureYield™ Plasmid Miniprep System" (Promega) dans 30 µl d'eau ultrapure.

Le dosage des plasmides purifiés a été réalisé à partir des valeurs d'absorbance à 260 nm déterminées par mesure entre 220 et 348 nm à l'aide d'un spectrophotomètre Varioskan (Thermo Scientific). Leur séquençage a été assuré par une entreprise tierce (GATC Biotech, Allemagne), à partir d'un mélange réactionnel de 10 µl comprenant 150 à 500 ng de plasmide et 25 pmol d'amorce Qe30_For dans de l'eau ultrapure. La validation des séquences a ensuite été réalisée après analyse des fichiers fournis au format FASTA et/ou des chromatogrammes de séquençage.

V.2.3.2. Sous-clonage des gènes codant pour les charpentes sauvages et greffées

Les gènes codant pour les séquences des charpentes protéiques humaines (structures 4F7H, 1LPJ, 1QNT) et de Sso7d (structure 1C8C) ont été synthétisés par MWG Eurofins, à partir de la séquence peptidique désirée entre les sites de restriction BamHI et HindIII, optimisée en acides nucléiques par le fournisseur pour une expression en souche *E. coli* K12 (**Tableau 7**). Des bactéries *E. coli* DH5α LacI^q ont été transformées avec 100 ng de chaque plasmide ainsi obtenu (procurant une résistance à l'ampicilline), puis les plasmides ont été amplifiés, purifiés et dosés comme décrit dans la section "V.2.3.1. Isolement, amplification et purification de plasmides". Après dosage, les plasmides ont été soumis à une double digestion par les enzymes de restrictions BamHI et HindIII, par incubation pendant 45 minutes à 37°C suivie d'une inactivation pendant 10 minutes à 80°C, dans le mélange réactionnel suivant:

Plasmide	1 µg
Tampon FD 10X	5 µl
BamHI	1 µl
HindIII	2 µl
Eau ultrapure	Qsp 50 µl

Après migration des échantillons par électrophorèse sur gel d'agarose à 1,5% pendant 45 minutes à 110 V, les inserts digérés ayant migré aux tailles attendues pour les gènes synthétiques ont été excisés et purifiés à l'aide du kit de dessalage "Wizard® SV Gel and PCR Clean-Up System" (Promega) et élués dans 30 µl d'eau ultrapure. Les inserts purifiés ont ensuite été clonés par ligation dans le vecteur (pAFG01, pAFG19, pAFG20, ou pAFG21) préalablement digéré par BamHI et HindIII en présence de phosphatase alcaline, par incubation pendant 30 minutes à température ambiante suivie d'une inactivation pendant 10 minutes à 65°C, dans le mélange réactionnel suivant:

Plasmide digéré BamHI/HindIII déphosphorylé	100 ng
Insert digéré BamHI/HindIII	3 excès molaires par rapport au plasmide
Tampon T4 10X	1 µl
T4 DNA ligase	0,5 µl
Eau ultrapure	Qsp 10 µl

La sélection des plasmides circulaires ayant intégré un insert a été réalisée par digestion par KpnI ou Sali, dont les sites uniques sont présents au niveau du site de multiclonaage des vecteurs entre les sites BamHI et HindIII, par incubation pendant 15 minutes à 37°C suivie d'une inactivation pendant 5 minutes à 80°C, dans le mélange réactionnel suivant:

Produit de ligation	10 µl
MgCl ₂ 50 mM	6 µl
KpnI ou Sali	1 µl
Eau ultrapure	23 µl

KpnI a été préférentiellement utilisée, hormis dans les cas où un site de restriction KpnI est présent dans la séquence codant pour la protéine à exprimer. Des bactéries *E. coli* DH5α LacI^q ont été transformées avec 5 µl du produit de ligation, puis les plasmides ont été amplifiés, purifiés et dosés comme décrit dans la section “V.2.3.1. Isolement, amplification et purification de plasmides”.

Tableau 7: Gènes synthétiques commandés pour le clonage des gènes des protéines humaines et de Sso7d.
WT: Sauvage. C6 et C8: Variants avec mutations par substitution pour greffer, respectivement, les feuillettes anti-GFP de C6 et C8.

Construction	Séquence du gène synthétique (5' vers 3')
1C8C(C6)	CATCACGGATCCGCCACCGTCAAATCAAGTATAAAGGCCAAGAGAAGAAGTGGACATTAGCAAGATTACCTTAGTGGTTCGCTTTGGGAA ACTGATCACGTTTCATTACGATGAAGGTGGGCAAAATGGGCACTGGAGCAGTACAGGAGAAGAATGCGCCGAAAGAAGCTTTGCAAATG CTGGAAGAGCAGAAAAAGCTTTATATGG
1C8C(C8)	CATCACGGATCCGCCACTGTCAAATCAAGTACAAGGCCAAGAGAAGAAGTGGACATTAGCAAGATTGGGCGTAGGGCGCTTAGGCA AAGAAATCACGTTTGGCTATGATGAAGGTGGAGGCAAAATGGGTGCTGGTCTGTACCAGAGAAGATGCGCCGAAAGAAGCTTGCAAAT GCTGGAAGAGCAGAAAAAGCTTTATATGG
4F7H(C6)	CATCACGGATCCATGACCAGCGAAAAATCACCTGAACAATTCGGACAAGAAGTGGACGAAGTTGATGCTGCGTGTCTGATCTCGAATTAC GCTGGAAGGTGGCAAAACATCCACCTTCTGGGCGATACCTTCGATTCGGAAATTAGCCGACTACATCAAAGTGTCAAACCGAAGAAGC TGACTCTGAAAGGTTACAAACAGTATTTGGTGTACGTTTAAGGATACAGTATTAGCTGCTACAAGTCAAAGAAGAGAGCTCTGGAACACCA GCACATCAGATGAACCTTCGTGGCACTTGTAGTCTCCGTTTGTGAACATCTCAGGTCAGAAATTCCTGATCAGCTTCACATTCCTGTAGCC GAAGGTTGGAATACCATTGCGCTGCAATGCGATAACGAGAACAGTATGCCATTGGATGGCTGCATGCTGCGCTGCAAAAGGCAAAAC CATGGCGGATAGTAGCTATAACCTGGAAGTCCAGAAATATCTGTCTTCTCAAATGCAAAAGCTTTATATGG
4F7H(C8)	CATCACGGATCCATGACGTCAGAAAAACACTTGAACAATTCGGATAAAGAGGTGGATGAAGTCGATGCCGATTGAGCGATCTCGAAATCAC GCTTGAAGGCGGTAACACGTTCTACCTTCTGGGCGCATTACAGCATTCCGGAATTAGCCGACTACATCAAGGTGTTAAACCGAAGAAGAAC TGACCTGAAAGGCTACAAACAGTACTGGTGTACCTTCAAAGACACCTCCATTTCGTGCTATAAGTCGAAAGAGGAATCGTCAGGTAACCA GCCATCAGATGAATCTCCGCGTTGGGCTGTTGGGCGTTAGTGAACATCAGTGGGCAAAAGTTGAGATCACACTCGGAATTCCTGTAGC AGAAGGTCGAATGCGATTGCGCTGACATGCGATAACGAGAACAGTATGCGATTGGATGGCTGCATGCTGCGGAGTAAAGGCAAAAC CTATGGCGGATAGCAGCTATAACCTGGAAGTTCAGAAATATCTGTCTTCTCAAATGCAAAAGCTTTATATGG
1LPJ(WT)	CATCACGGATCCATGCCAGCTGATCTTTTCAGGCACATGGACCTTGTAAAGCAGTGACAACCTTGAAGGTTATATGCTGGCACTGGGCATTGA CTTTGCGACGCGTAAATTCGCCAACTGCTGAAACCGCAGAAAGTGAACGACAGATGCGGATTCGTTTACCAATTCACCAATTCAGCTT ACGCAACTACTTCGTCGAAGTTCAAAGTAGGGGAAGAATTCGATGAGGATAATCGCGGTTTGGACAATCGCAAATGCAAAATCCTCGTTATCT GGGATAACGATCGTCTGACCTGTATTAGAAAGGAGAGAAGAAGAACCGTGGTGGACTCATTGGATTGAAGGCGCAAACTGCATCTGGA GATGTTTTGTGAAGTCAAGTGTGCAAAACAGACGTTTCAACGGGCAAGCTTTATATGG
1LPJ(C6)	CATCACGGATCCATGCCAGCAGACTAAGTGGGACATGGACGCTGTGTCATCCGACAACCTTGAAGGCTATATGCTGGCCTTAGGTATCGA CTTTGCTACTCGCAAATTCGCAAACTCTGAAACCGCAGAAAGTGAATGAGCAGAAATGGCGATAGCTTTACCAATTCATCAATAGCTCTTT GCCAACTACTTCGTGAAGTTCAAAGTTGGGCAAGAATTCGATGAGGACAATCGTGGTTCGGATAACCGCAAATGCAAAATCCTCGTTAATCT GGGATAACGATCGTCTGACGTCATTCAGAAAGGCGAAAGAAGAATCGTGGAACTCTGCATGTCATCTTTGGGGATCTCTGACCTCCAC

	ATGTTCTGTGAAGGTCAGTGGTGTACCCAAGCGTTTCAACGGGGCAAGCTTTATATGG
1LPJ(C8)	CATCACGGATCCATGGCAGCAGATCTGTCCAGAACCTGGACACTGCTCAGCAGTGACAACCTCGAAGGCTATATGCTGGCCTTAGGCATCGA CTTTGGCGACTCGCAAGATTGCCAACTGCTGAAACCGCAGAAAAGTGATTGAGCAGAAATGTTGATTCCTTTACCATCCACACCAATAGCTCTTT GGCACAATACTTCGTCAAATTCAAAGTAGGCGAAGAGTTTGGACGAAGATAATCGTGGTCTGGATAACCGCAAATGCAATCGCTGTTATCT GGGATAATGACCGTCTTACGTGCATTAGAAAAGCGAAAAGAAGAACCGTGGTGGCTCATGGGATTTGGGCGATGAAGTACCTTAGG GATGTTTTGTGAAGGTCAACTGTGTGCGCAACGGTTTACGCGCGCAAGCTTTATATGG
1QNT(WT)	CATCACGGATCCATGGACAAGGATTGCGAGATGAAACGCACAACTGGATTCTCCGCTTGGCAAACCTGGAACCTCGGGTTCGCAACAGGG CCTGCACGAAATCAAAGTCTGGGTAAGGCACGAGTGGCGGTGATGCCGTCGAAGTTCCTGCACACCGCCGCTTGGGAGGTCCGGAAC CGCTCATGCGAGTGTACTGCGTGGTTAAACGCGTATTTCCACCAACCGGAAGCGATTGAGGAGTTTCCGTTCGCCCTTACACCATCCGGTGT TTCAGCAGGAATCGTTACCCGTCAGTGTCTGGAACTGCTGAAAGTGGTGAAGTTTGGGGAAGTATTAGTACCAGCAATGGCCGCA TTGGCGGGCAATCCGAAAGCAGCTCGTGGGTTGGCGGTGCAATGCGCGGCAATCCCGTCCCAATCCTGATTCCGTGTCATCGGGTAGTTG CTCTCAGGTGCGGTAGGCAACTATAGCGGAGGGTTAGCCGTGAAAGAGTGGCTGCTGGCTCATGAAGTTCATCGCTTAAGCTTTATATGG
1QNT(C6)	CATCACGGATCCATGGACAAGGATTGCGAGTGGAAAACCCGCGTTACAGAGCCGCTTGGCTTACTTACCCTGCATGGTTGCGAACAAGG GCTGACGCTCATTGTCTGTTCGGCAAAGGCACGTCAGCGGTGATGCCGTAGAAGTTCGGCGCCAGCCGAGTCTCGTGGTCTCTGAAC CCCTGATGCGAGTGTACAGCGTGGCTGAACGCGTACTTTCCACCAACCGGAAGCCATCGAGGAATTTCCGGTCCAGCCTTGCATCACCCCTGTG TTTACAGCAGGAATCGTTCACTCGCAAGTGTGTGGAACTGCTGAAAGTGGTCAAGTTCGGTGAAGTATCAGCAGTATCAGCAGCTGGCAGC ATTGGCCGGCAATCCCAAAGCGCTCGTCCGTTGGTGGAGCTATGCGTGGCAATCCGGTCCGATTCTGATTCCGTGTCATCGCGTAGTTT GCAGTCTGGCCAGTAGGGAATTTCCGGTGGATTGGCGGTCAAAGAGTGGCTGTTAGCGCACGAAGGCCATCGCTGAAGCTTTATATGG
1QNT(C8)	CATCACGGATCCATGGACAAGGATTGCGAGCTTAAAGCCACACGCTGACTCTCCGTTAGGCGAACTGACCCTTGGCGGGTGTGAGCAGGG CTTGTGGGCCATTGGCTGCTGGGAAAAGGCACCTCAGCGGTGATGCGAGTGAAGTTCCTGCACACCGGCTGCTTGGGCGTCCGGA CCGCTGATGCGAGTGTACGGCGTGGTTGAACGCGTACTTCCACCAACCGAAGCGATTGAGGAATTTCCGGTTCCTGCGCTGCATCACCCCTGTG ATTTACAGCAGGAGTGTGTTACCGCGCAAGTGTGTGGAACTCCTGAAAGTGGTCAAATTCGGGGAAGTATCTCTATCAGCAACTCGCCG CCTTAGCGGCAATCCCAAAGTCCCGTGGCGGTGGGAGGTGCAATGCGCGGTAATCCGGTTCGGATCCTGATTCCGTGCCATCGGCTGGTT TGCAGCAGTGGTGCAGTAGGGAATATAGCGTGGTCTGGCAGTCAAGGAATGGCTGCGCATGAAGGCCACCGCTTAAAGCTTTATAT TGG

V.2.3.3. Construction des plasmides pAFG12, pAFG19, pAFG20, pAFG21

Les gènes codant pour les protéines GFP, TrxA, NusA, et MBP ont été clonés dans le vecteur pAFG16 par digestion et ligation entre les sites de restriction EcoRI et BamHI, aboutissant à des constructions composées de la protéine fusionnée en amont du site BamHI et d'une étiquette RGS-His₆ en aval du site HindIII. Les gènes des protéines de fusion ont été amplifiés par PCR entre les amorces **{Fusion}_E1_For** et **{Fusion}_B1_Rev**, où **{Fusion}** représente le nom de la fusion (GFP, Trx, NusA ou MBP), en utilisant les mélanges réactionnels et cycles suivants (à l'aide d'un thermocycleur T100, BioRad):

Composition du mélange réactionnel, en µl:

Programme PCR "FUSIONS>ALL":

Eau ultrapure	37	98°C	30 sec	} 30X
Phusion 5X buffer HF	10	98°C	10 sec	
dNTP mix 10 mM	1	60°C	30 sec	
Amorce {Fusion}_E1_For 10 µM	0,5	72°C	90 sec	
Amorce {Fusion}_B1_Rev 10 µM	0,5	72°C	5 min	
Phusion DNA polymerase	0,5	14°C	∞	
Matrice	0,5			
Volume final	50			

Les matrices employées pour construire les vecteurs pAFG12, pAFG19, pAFG20 et pAFG21 (correspondant respectivement aux fusions GFP, TrxA, NusA et MBP) sont, respectivement, les vecteurs pAFG01-GFP, pNH-TRXT, pET-His6-NusA-TEV-LIC, pRK793. Les trois derniers vecteurs ont été aimablement fournis par Michel Dion. Un témoin négatif a été réalisé pour chaque réaction de PCR, dans les mêmes conditions en remplaçant la matrice par de l'eau ultrapure. Après analyse par électrophorèse sur gel d'agarose à 1,5% pendant 45 minutes à 110 V, les amplicons ont

été purifiés à l'aide du kit de dessalage “Wizard® SV Gel and PCR Clean-Up System” (Promega) et élués dans 30 µl d'eau ultrapure. Après dosage, les inserts ont été soumis à une double digestion par les enzymes de restrictions BamHI et EcoRI, par incubation pendant 45 minutes à 37°C dans le mélange réactionnel suivant:

Insert	2 µg
Tampon FD 10X	10 µl
BamHI	2 µl
EcoRI	4 µl
Eau ultrapure	Qsp 100 µl

Les inserts digérés ont été immédiatement purifiés à l'aide du kit de dessalage “Wizard® SV Gel and PCR Clean-Up System” (Promega) et élués dans 30 µl d'eau ultrapure, puis clonés par ligation dans le vecteur pAFG16 préalablement digéré par BamHI et EcoRI en présence de phosphatase alcaline, selon la méthode décrite dans la section “V.2.3.2. Sous-clonage des gènes codant pour les charpentes sauvages et greffées” (jusqu'à validation des plasmides par séquençage).

V.2.3.4. Construction de Nanofitines ponctuelles (NF1 à NF10, variants alanine de C8)

Les gènes codant pour les Nanofitines non générées par sélection *in vitro* ont été assemblés de façon similaire à la banque dont la construction est décrite dans la section “V.2.2.2. Tours de sélection par *ribosome display* et isolement des clones”, selon la méthode suivante.

En résumé, les gènes ont été assemblés par deux réactions de PCR chevauchantes successives (**Figure 80**). La première code pour les régions normalement randomisées dans la banque, à partir d'oligonucléotides “2.2”, “2.3” et “2.4”. La seconde ajoute les régions constantes des Nanofitines ainsi que les sites de restriction BamHI et HindIII, respectivement aux extrémités 5' et 3'.

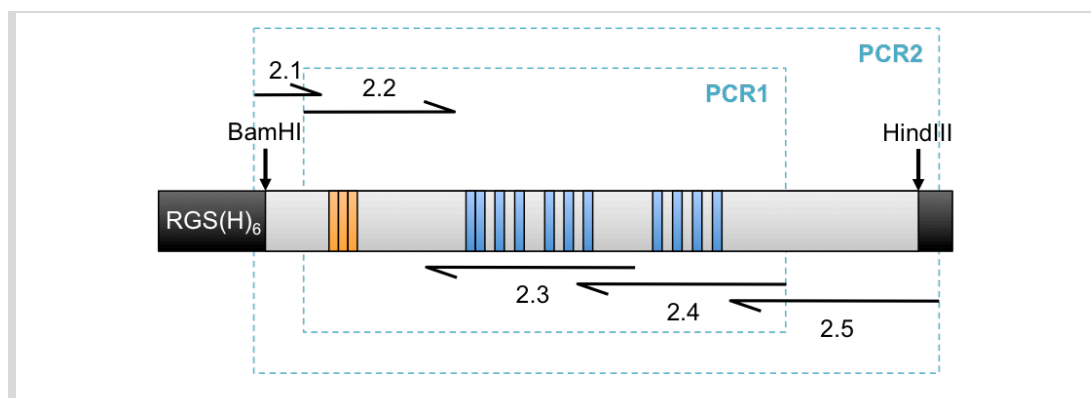


Figure 80 : Représentation schématique de l'assemblage d'une séquence de Nanofitine ponctuelle par PCR chevauchantes. La portion grise correspond à la séquence de *Sac7d* comprise entre les sites de restriction BamHI et HindIII. Les mutations dans la première boucle et à la surface du second feuillet beta introduites pour générer des Nanofitines sont colorées, respectivement, en orange et en bleu. Les portions figurées en noir correspondent aux extensions de séquence ajoutées lors des assemblages par biologie moléculaire. Les amorces sens (2.1 et 2.2) et antisens (2.3, 2.4 et 2.5) sont représentées colinéairement à la séquence des Nanofitines.

V.2.3.4.1. PCR1: Assemblage du cœur variable

Les régions variables des Nanofitines ont été amplifiées par PCR entre les amorces “2.2”, “2.3” et “2.4”, selon les combinaisons listées dans le **Tableau 8**, en utilisant les mélanges réactionnels et cycles suivants (à l’aide d’un thermocycleur T100, BioRad):

<u>Composition du mélange réactionnel, en µl:</u>		<u>Programme PCR “PCR NNS”:</u>	
Eau ultrapure	35,5	98°C	30 sec
Phusion 5X buffer HF	10	98°C	30 sec
DMSO	1,5	50°C	30 sec
dNTP mix 10 mM	1	72°C	30 sec
Amorce “2.2” 10 µM	0,5	98°C	30 sec
Amorce “2.3” 10 µM	0,5	64°C	30 sec
Amorce “2.4” 10 µM	0,5	72°C	30 sec
Phusion DNA polymerase	0,5	72°C	5 min
Volume final	50	14°C	∞

} 5X
 } 30X

Un témoin négatif a été réalisé dans les mêmes conditions en remplaçant les amorces par de l’eau ultrapure. Après analyse par électrophorèse sur gel d’agarose à 1,5% pendant 45 minutes à 110 V, les amplicons ont été purifiés à l’aide du kit de dessalage “Wizard® SV Gel and PCR Clean-Up System” (Promega) et élués dans 30 µl d’eau ultrapure.

Tableau 8: Combinaisons des oligonucléotides utilisés pour la construction de la région variable des Nanofitines.

Nanofitines	Amorce “2.2” sens	Amorce “2.3” antisens	Amorce “2.4” antisens
Variants alanine de C8 anti-GFP	SclibKYK.2_for	1208_C8_2.3_rev	1208_C8_Ala8_rev 1208_C8_Ala10_rev 1208_C8_Ala11_rev
		1208_C8_Ala1_rev 1208_C8_Ala3_rev 1208_C8_Ala4_rev 1208_C8_Ala5_rev 1208_C8_Ala6_rev 1208_C8_Ala7_rev	1208_C8_2.4_rev
NF1 à NF5	NF2.2_VFK NF2.2_VFR	NF2.3_KTDHITS NF2.3_KTSHITS	NF2.4_EAAE NF2.4_EAAH
NF6 à NF9	NF2.2_VHD	NF2.3_KGWHVAS NF2.3_KAWWIAS NF2.3_KGWHWHS	NF2.4_VERH NF2.4_VESH
NF9 et NF10	NF2.2_VFN	NF2.3_FLDMMWWT	NF2.4_WRYH NF2.4_WRFH

V.2.3.4.2. PCR2: Assemblage des régions constantes

Les régions constantes des Nanofitines ont été ajoutées par PCR grâce aux amorces “2.1” et “2.5”, en utilisant les mélanges réactionnels et cycles suivants (à l’aide d’un thermocycleur T100, BioRad):

<u>Composition du mélange réactionnel, en µl:</u>		<u>Programme PCR “D1SSOB”:</u>	
Eau ultrapure	Qsp 50	98°C	30 sec
Phusion 5X buffer HF	10	98°C	30 sec
DMSO	1,5	64°C	30 sec
dNTP mix 10 mM	1	72°C	30 sec
Amorce Sclib2.1_For 10 µM	0,5	72°C	5 min
Amorce Sclib2.5_Rev 10 µM	0,5	14°C	∞
Phusion DNA polymerase	0,5		
Produit de PCR1	100 ng		
Volume final	50		

20X

Un témoin négatif a été réalisé pour chaque réaction de PCR, dans les mêmes conditions en remplaçant la matrice par de l’eau ultrapure. Après analyse par électrophorèse sur gel d’agarose à 1,5% pendant 45 minutes à 110 V, les amplicons ont été purifiés à l’aide du kit de dessalage “Wizard® SV Gel and PCR Clean-Up System” (Promega) et élués dans 30 µl d’eau ultrapure. Après dosage, les amplicons ont été soumis à une double digestion par les enzymes de restrictions BamHI et HindIII, pour être clonés dans le plasmide pAFG01, de façon identique à celle décrite dans la section “V.2.3.2. Sous-clonage des gènes codant pour les charpentes sauvages et greffées”.

V.2.3.5. Changement de résistance du plasmide pAFG05-GFP

Pour pouvoir co-exprimer la StrepTagII-GFP avec des protéines fusionnées à une étiquette RGS-His₆ (exprimées à l’aide des plasmides pAFG01, pAFG19, pAFG20 ou pAFG21), le gène de résistance au chloramphénicol a été substitué à celui de résistance à l’ampicilline dans le plasmide pAFG05-GFP.

La séquence du promoteur et du gène de résistance au chloramphénicol a été amplifiée entre les sites de restriction PstI et KpnI par PCR avec les amorces CamR_PstI_for et CamR_KpnI_rev, en utilisant les mélanges réactionnels et cycles suivants (à l’aide d’un thermocycleur T100, BioRad):

<u>Composition du mélange réactionnel, en µl:</u>		<u>Programme PCR “FUSIONS>ALL”:</u>	
Eau ultrapure	37	98°C	30 sec
Phusion 5X buffer HF	10	98°C	10 sec
dNTP mix 10 mM	1	60°C	30 sec
Amorce CamR_PstI_for 10 µM	0,5	72°C	90 sec
Amorce CamR_KpnI_rev 10 µM	0,5	72°C	5 min
Phusion DNA polymerase	0,5	14°C	∞
Matrice (pbirAcm à ~200ng/µl)	0,5		
Volume final	50		

30X

La séquence du plasmide pAFG05-GFP a été amplifiée entre les sites de restriction PstI et KpnI (sans la séquence du promoteur et du gène de résistance à l'ampicilline) par PCR avec les amorces AmpRdel_PstI_for et AmpRdel_KpnI_rev, en utilisant les mélanges réactionnels et cycles suivants (à l'aide d'un thermocycleur T100, BioRad):

<u>Composition du mélange réactionnel, en µl:</u>		<u>Programme PCR "GIBS-PL":</u>	
Eau ultrapure	37	98°C	30 sec
Phusion 5X buffer HF	10	98°C	10 sec
dNTP mix 10 mM	1	60°C	30 sec
Amorce AmpRdel_PstI_for 10 µM	0,5	72°C	90 sec
Amorce AmpRdel_KpnI_rev 10 µM	0,5	72°C	5 min
Phusion DNA polymerase	0,5	14°C	∞
Matrice	0,5		
Volume final	50		

Un témoin négatif a été réalisé pour chaque réaction de PCR, dans les mêmes conditions en remplaçant la matrice par de l'eau ultrapure. Après analyse par électrophorèse sur gel d'agarose à 1,5% (pour le gène de résistance au chloramphénicol) et 0,8% (pour le plasmide) pendant 45 minutes à 110 V, les amplicons ont été purifiés à l'aide du kit de dessalage "Wizard® SV Gel and PCR Clean-Up System" (Promega) et élués dans 30 µl d'eau ultrapure. Après dosage, les amplicons ont été soumis à une double digestion par les enzymes de restrictions PstI et KpnI, par incubation pendant 45 minutes à 37°C dans les mélanges réactionnels suivants:

<u>Double-digestion de l'insert:</u>		<u>Double-digestion du plasmide:</u>	
Insert	2 µg	Insert	2 µg
Tampon FD 10X	10 µl	Tampon FD 10X	10 µl
PstI	4 µl	PstI	4 µl
KpnI	2 µl	KpnI	2 µl
Eau ultrapure	Qsp 100 µl	Phosphatase alcaline	2 µl
		Eau ultrapure	Qsp 100 µl

Les amplicons digérés ont été immédiatement purifiés à l'aide du kit de dessalage "Wizard® SV Gel and PCR Clean-Up System" (Promega) et élués dans 30 µl d'eau ultrapure, puis assemblés par ligation, selon la méthode décrite dans la section "V.2.3.2. Sous-clonage des gènes codant pour les charpentes sauvages et greffées" (jusqu'à validation des plasmides par séquençage) en remplaçant l'ampicilline par du chloramphénicol à une concentration finale de 10 µg/ml.

V.2.3.6. Construction de tétramères de Nanofitine

Les gènes codant pour les Nanofitines tétramériques ont été construits et clonés dans le vecteur pAFG01 en une seule étape par assemblage Gibson (*Gibson et al., 2010*), aboutissant à des constructions composées de 4 sous-unités (de l'extrémité N-terminale à l'extrémité C-terminale des protéines codées) avec étiquette RGS-His₆ en fusion N-terminale. Le vecteur de destination a

été amplifié par PCR entre les amorces Gpls01C_For et Gpls01N_Rev, en utilisant les mélanges réactionnels et cycles suivants (à l'aide d'un thermocycleur T100, BioRad):

<u>Composition du mélange réactionnel, en µl:</u>		<u>Programme PCR "GIBS-PL":</u>	
Eau ultrapure	37	98°C	30 sec
Phusion 5X buffer HF	10	98°C	10 sec
dNTP mix 10 mM	1	60°C	30 sec
Amorce sens 10 µM	0,5	72°C	90 sec
Amorce antisens 10 µM	0,5	72°C	5 min
Phusion DNA polymerase	0,5	14°C	∞
Matrice	0,5		
Volume final	50		

} 25X

Les séquences des Nanofitines à multimériser ont été amplifiées par PCR entre une amorce sens et une amorce antisens en fonction de leur position dans le tétramère, comme résumé dans le **Tableau 9**, en utilisant les mélanges réactionnels et cycles suivants (à l'aide d'un thermocycleur T100, BioRad):

<u>Composition du mélange réactionnel, en µl:</u>		<u>Programme PCR "INSERT":</u>	
Eau ultrapure	37	98°C	30 sec
Phusion 5X buffer HF	10	98°C	10 sec
dNTP mix 10 mM each	1	60°C	30 sec
Amorce sens 10 µM	0,5	72°C	30 sec
Amorce antisens 10 µM	0,5	72°C	5 min
Phusion DNA polymerase	0,5	14°C	∞
Matrice	0,5		
Volume final	50		

} 25X

Tableau 9: Oligonucléotides utilisés pour la construction de tétramères.

	1ère sous-unité	2ème sous-unité	3ème sous-unité	4ème sous-unité
Amorce sens	Gpls01N_For	Glink5.1_For	Glink5.2_For	Glink5.3_For
Amorce antisens	Glink5.1_Rev	Glink5.2_Rev	Glink5.3_Rev	Gpls01C_Rev

Un témoin négatif a été réalisé pour chaque réaction de PCR, dans les mêmes conditions en remplaçant la matrice par de l'eau ultrapure. Après analyse par électrophorèse sur gel d'agarose à 0,8% (pour le plasmide) et 1,5% (pour les sous-unités) pendant 45 minutes à 110 V, les amplicons ont été purifiés à l'aide du kit de dessalage "Wizard® SV Gel and PCR Clean-Up System" (Promega) et élués dans 30 µl d'eau ultrapure.

L'assemblage des fragments a été réalisé en mélangeant le plasmide linéarisé (100 ng) avec 5 excès molaires de chacun des inserts, dans un volume final de 5 µl. Ensuite, 15 µl de mélange réactionnel d'assemblage Gibson (25% PEG-8000, 500 mM Tris-HCl, 50 mM MgCl₂, 50 mM DTT, 1 mM Mix dNTPs, 5 mM NAD, 2U de T5 exonuclease, 12.5U de Phusion polymerase, 2000U de Taq

ligase) ont été ajoutés et la solution a été incubée pendant 1 heure à 50°C. Des bactéries *E. coli* DH5 α LacI^q ont finalement été transformées avec 10 μ l de mélange réactionnel, puis les plasmides ont été purifiés et caractérisés comme décrit dans la section “V.2.3.1. Isolement, amplification et purification de plasmides”.

Les plasmides s'étant soldés par des constructions avec des sous-unités manquantes (dimères ou trimères) mais constituant un cadre de lecture correct ont également été utilisés pour l'expression bactérienne et la purification.

V.2.3.7. Construction de Nanofitine sans étiquette de fusion

Le gène codant pour la Nanofitine NF5 a été construit et cloné dans un vecteur dérivé de pAFG01 par assemblage Gibson de façon similaire à celle décrite dans la section “V.2.3.6. Construction de tétramères de Nanofitine”, aboutissant à une construction sans étiquette RGS-His₆ en fusion N-terminale. La différence réside dans le choix des oligonucléotides utilisés: le vecteur de destination a été amplifié par PCR entre les amorces Gpls01C_For et Gpls02N_RevGS avec le programme PCR “GIBS-PL”, tandis que la séquence codante de la Nanofitine a été amplifiée par PCR entre les amorces Gpls02N_ForGS et Gdel3AA_Rev avec le programme PCR “INSERT”. L'assemblage Gibson a été effectué de la même façon, avec 100 ng de plasmide linéarisé et 5 excès molaire de l'insert de Nanofitine.

V.2.3.8. Construction dédiées à la cristallogénèse

V.2.3.8.1. Plasmide pAFG01-GFP Δ linker

Pour augmenter les rendements de purification par chromatographie d'affinité en vue de la cristallogénèse de la GFP, son gène codant a été cloné à l'extrémité 3' de la séquence de l'étiquette RGS-His₆. L'espaceur flexible entre l'étiquette polyhistidine et la GFP a également été réduit pour limiter de potentiels mouvements inhibiteurs de cristallogénèse, aboutissant à la séquence N-terminale MRGSHHHHHHGS, suivie de la méthionine initiale de la GFP. Cette fusion se distingue donc de celles codées par les vecteurs pAFG01-GFP ou pAFG05-GFP préalablement générés à Affilogic, ajoutant respectivement les séquences MRGSHHHHHHGSACELGTLGSAGSAAGSGEF ou MASWSHPQFEKGSAGSAAGSGEF avant la méthionine initiale de la GFP.

Le gène codant pour la GFP a été cloné dans un vecteur dérivé de pAFG01 par assemblage Gibson de façon similaire à celle décrite dans la section “V.2.3.6. Construction de tétramères de Nanofitine”, aboutissant au vecteur pAFG01-GFP Δ linker. La différence réside dans le choix des oligonucléotides utilisés: le vecteur de destination (pAFG01) a été amplifié par PCR entre les amorces Gpls03C_For et Gpls01N_Rev avec le programme PCR “GIBS-PL”, tandis que la séquence codante de la GFP a été amplifiée par PCR à partir de pAFG05-GFP entre les amorces HisGFPN_For et Gpls03C_Rev avec le programme PCR “INSERT”. L'assemblage Gibson a été effectué de la même façon, avec 100 ng de plasmide linéarisé et 5 excès molaire de l'insert de GFP.

V.2.3.8.2. Chimères GFP-NF5 et GFP-C8

Pour s'assurer d'obtenir un ratio GFP:Nanofitine de 1:1 lors du criblage des conditions de cristallogénèse des complexes entre la GFP et NF5 ou C8, des vecteurs d'expression ont été construits pour fusionner l'étiquette RGS-His₆, la GFP sans espaceur superflux (GFP Δ linker), un bras espaceur de longueur variable, puis la Nanofitine (de l'extrémité N-terminale à l'extrémité C-terminale). D'après les structures prédites de complexe GFP:NF5, une fixation intramoléculaire serait possible avec un bras espaceur de courte taille (≤ 10 résidus) entre l'extrémité C-terminale de la GFP et l'extrémité N-terminale de la Nanofitine. Pour favoriser des interactions intramoléculaires tout en permettant la formation d'interactions intermoléculaires, des constructions avec des espaceurs flexibles (composés de glycines, sérines et alanines) de 7 (L7), 10 (L10) et 20 (L20) résidus ont été réalisés. Les 3 derniers résidus des Nanofitines (KLN), suspectés d'être clivés d'après les données d'autres projets, ont été jugés comme possible source d'hétérogénéité et ont également été supprimés.

Les gènes codant pour la GFP et les Nanofitines ont été clonés dans un vecteur dérivé de pAFG01 par assemblage Gibson de façon similaire à celle décrite dans la section "V.2.3.6. Construction de tétramères de Nanofitine", aboutissant aux vecteurs pAFG01-GFP-L7-NF5, pAFG01-GFP-L7-C8, pAFG01-GFP-L10-NF5, pAFG01-GFP-L10-C8, pAFG01-GFP-L20-NF5 et pAFG01-GFP-L20-C8. La différence réside dans le choix des oligonucléotides utilisés: le vecteur de destination (pAFG01) a été amplifié par PCR entre les amorces Gpls01_For et Gpls01N_Rev avec le programme PCR "GIBS-PL"; la séquence codante de la GFP a été amplifiée par PCR à partir de pAFG05-GFP entre les amorces HisGFPN_For et GFPlink{espaceur}_Rev avec le programme PCR "INSERT"; la séquence codante de la Nanofitine a été amplifiée par PCR à partir du plasmide pAFG01 (possédant l'insert codant NF5 ou C8) entre les amorces GFPlink{espaceur}_For et HisGFPN_For et Gdel3AA_Rev avec le programme PCR "INSERT". Des oligonucléotides GFPlink{espaceur}_Rev et GFPlink{espaceur}_Rev différents ont été utilisés, avec {espaceur} correspondant à 5.1, 10.3 ou 20.1, pour générer respectivement les espaceurs flexibles L7 (GGGSAGS), L10 (GGAGGSGAGS) ou L20 (GGSAGSSGASGGAGSGGS). L'assemblage Gibson a été effectué de la même façon, avec 100 ng de plasmide linéarisé et 5 excès molaire des inserts de GFP et de Nanofitine.

V.2.3.9. Mutagenèse en alanine des variants de la GFP

Les plasmides des GFP avec mutations ponctuelles en alanine ont été générés à partir du plasmide pAFG05-GFP (codant pour la protéine StrepTagII-GFP) par PCR selon la méthode *Quick-change* (Agilent-Technologies, 1991). Chaque mutation a été introduite grâce à un couple d'oligonucléotides sens et antisens incorporant la mutation désirée, respectivement nommés AcGFP-{mutation}_For et AcGFP-{mutation}_Rev, où {mutation} représente la substitution incorporée (ex: E95A). Les réactions de PCR ont été réalisées en utilisant les mélanges réactionnels et cycles suivants (à l'aide d'un thermocycleur T100, BioRad):

Composition du mélange réactionnel, en μ l:		Programme PCR "QuickGFP":		
Eau ultrapure	Qsp 50	95°C	30 sec	} 18X
10X Pfu buffer + MgSO ₄	5	95°C	30 sec	
dNTP mix 10 mM	1	55°C	1 min	
Amorce sens 10 μ M	0,5	68°C	4 min 30 sec	
Amorce antisens 10 μ M	0,5	14°C	∞	
Pfu DNA polymerase	1			
Matrice (pAFG05-GFP)	10 ng			

Un témoin négatif a été réalisé pour chaque réaction de PCR, dans les mêmes conditions en remplaçant la matrice par de l'eau ultrapure. Après analyse par électrophorèse sur gel d'agarose à 1% pendant 45 minutes à 110 V, 1 μ l d'enzyme de restriction DpnI (DpnI FastDigest, Fermentas) a été ajouté aux mélanges réactionnels pour cliver les brins d'ADN non mutés pendant une incubation à 37°C pendant 90 minutes, puis les amplicons ont été purifiés à l'aide du kit de dessalage "Wizard® SV Gel and PCR Clean-Up System" (Promega) et élués dans 30 μ l d'eau ultrapure. Des bactéries *E. coli* DH5 α LacI^q ont finalement été transformées avec 100 ng de produit purifié, puis les plasmides ont été isolés, purifiés et caractérisés comme décrit dans la section "V.2.3.1. Isolement, amplification et purification de plasmides".

V.2.4. Expression et purification des protéines recombinantes

V.2.4.1. Cultures bactériennes

Les Nanofitines, les variants de GFP, les charpentes alternatives et les constructions chimériques fusionnées avec les protéines TrxA, NusA, MBP ou GFP ont été exprimés en bactéries *E. coli* DH5 α LacI^q. En résumé, des précultures ont été inoculées avec une carotte de souche en glycérol, puis incubée pendant à 37°C pendant la nuit en milieu 2xYT additionné de 100 μ g/ml d'ampicilline, de 25 μ g/ml de kanamycine et de 1% de glucose. Les précultures ont été diluées au 1:20 en milieu 2xYT additionné de 100 μ g/ml d'ampicilline, de 25 μ g/ml de kanamycine et de 0,1% de glucose, puis incubées à 37°C jusqu'à atteindre la phase exponentielle de croissance (absorbance à 600 nm comprise entre 0,8 et 1,0). L'expression des protéines recombinantes a ensuite été induite par ajout d'Isopropyl β -D-1-thiogalactopyranoside (IPTG) à une concentration finale de 0,5 mM et incubation à 30°C pendant la nuit. L'arrêt des cultures a été réalisé par centrifugation pendant 45 minutes à 3220g ou 20 minutes à 6000 rpm. Les surnageants de culture ont été éliminés puis les culots bactériens ont été lysés immédiatement ou conservés à -80°C avant lyse.

V.2.4.2. Cultures bactériennes alternatives

Des conditions alternatives ont été réalisées pour les cultures bactériennes des charpentes protéiques greffées des résidus fonctionnels des feuillets anti-GFP de C6 et C8. Les étapes de l'arrêt des cultures à la purification par chromatographie d'affinité à suivre ont été réalisées à 4°C, et non à température ambiante, avec ou sans ajout de lysozyme à 1 mg/ml (Fluka) lors de l'étape de lyse.

V.2.4.2.1. Modification des conditions d'induction

Des cultures ont été réalisées de la même façon que décrite précédemment, en abaissant la température d'induction à 20°C au lieu de 30°C. Des cultures ont également été réalisées sans induction en remplaçant le milieu 2xYT par du milieu auto-inductible ZYM5052 (ForMedium; Studier, 2005). Dans ce dernier cas, la température de culture a été abaissée de 37°C à 20°C après 3 heures de culture. La moitié du volume de culture a été collecté après 3,5 heures de culture, et la moitié restante a été collectée après incubation pendant la nuit, puis la lyse et la purification de chaque demi-culture bactérienne a été réalisée séparément.

V.2.4.2.2. E. coli de souche B avec chaperonnes

Alternativement, le protocole de culture a été réalisé à partir de plasmides transformés en bactéries *E. coli* SHuffle T7 Express lysY (New England Biolabs; de Marco, 2009), de la façon décrite dans la section "V.2.3.1. Isolement, amplification et purification de plasmides", pour permettre l'expression des protéines humaines insolubles, en remplaçant l'ampicilline par du chloramphénicol à une concentration finale de 10 µg/ml.

V.2.4.2.3. Co-expression StrepTagII-GFP et charpente greffée

Lors des conditions de cultures avec co-expression, du chloramphénicol a été ajouté à une concentration finale de 10 µg/ml lors de chaque étape nécessitant l'addition d'ampicilline. En résumé, des bactéries *E. coli* DH5α LacI^q exprimant les charpentes humaines ou Sso7d greffées des feuillets anti-GFP de C6 ou C8 (clonées en plasmide pAFG01) ont été transformées avec le plasmide dérivé de pAFG05-GFP (permettant l'expression de la StrepTagII-GFP) procurant la résistance au chloramphénicol, de la façon décrite dans la section "V.2.3.1. Isolement, amplification et purification de plasmides". Des colonies bactériennes fluorescentes isolées sur géloses de milieu 2xYT additionné de 100 µg/ml d'ampicilline, 25 µg/ml de kanamycine et 10 µg/ml de chloramphénicol ont ensuite été repiquées et inoculées pour réaliser des précultures et cultures à suivre dans les conditions standards préalablement décrites, avec addition de chloramphénicol aux milieux de culture. Lors de la co-purification de complexes entre protéines recombinantes anti-GFP (fusionnées à une étiquette poly-histidine) et la StrepTagII-GFP, la moitié de chaque culot bactérien a été traitée pour purification par IMAC et l'autre moitié a été traitée pour purification sur résine de Strep-Tactine, tel que décrit dans la section "V.2.4.3. Purification".

V.2.4.3. Purification

V.2.4.3.1. Protéines avec étiquette poly-histidine (IMAC)

Les culots bactériens collectés lors de l'arrêt des cultures ont été resuspendus dans 5 ml de tampon de lyse à pH 7,4 (composé de réactif d'extraction protéique BugBuster 1X, 5 µg/ml de DNaseI, 20 mM de Tris, 500 mM de NaCl, et 25 mM d'imidazole) par gramme de culot. La lyse a été réalisée par incubation pendant 1 heure à température ambiante sous agitation, puis la

suspension a été centrifugée à 3220g pendant 45 minutes pour éliminer les débris cellulaires (culot, ou fraction C).

La purification des protéines recombinantes fusionnées à une étiquette poly-histidine a ensuite été réalisée à partir du surnageant de lyse, par chromatographie d'affinité sur métal immobilisé (IMAC) à l'aide d'une résine His60 Nickel Superflow (Clontech). En bref, les résines ont été lavées par passage de 5 volumes d'eau ultrapure, puis équilibrées par passage de 5 volumes de tampon A (20 mM de Tris, 500 mM de NaCl, et 25 mM d'imidazole, pH 7,4). Le surnageant de lyse a été appliqué sur la résine par gravité, permettant de récolter la fraction non retenue (NR). Le lavage des résines ayant capturé les protéines recombinantes a été réalisé par passage de 30 volumes de tampon A. Finalement, l'élution a été effectuée par passage d'un volume de tampon B (20 mM de Tris, 500 mM de NaCl, et 250 mM d'imidazole, pH 7,4) pour éliminer le volume mort de résine, puis deux passages successifs de 2,5 volumes de tampon B (respectivement fractions d'élution EL1 et EL2). Des cycles supplémentaires de purification ont été réalisés après équilibrage de la résine par passage de 30 volumes de la fraction de lavage du cycle précédent, à partir de la fraction non retenue.

V.2.4.3.2. Protéines avec étiquette StrepTagII (Strep-Tactine)

Pour les variants de StrepTagII-GFP, le tampon de lyse a été réalisé à partir d'une solution à pH 7,4 composée de BugBuster 1X, 5 µg/ml de DNaseI, 20 mM de Tris, 150 mM de NaCl. La purification a été effectuée par chromatographie d'affinité à l'aide d'une résine Strep-Tactine Sepharose (IBA) équilibrée avec 20 ml de tampon W (composé de 100 mM de Tris et 150 mM de NaCl, pH 8). Après passage des surnageants de lyse, les lavages ont été assurés par 5 passages successifs de 2 volumes de tampon W, puis l'élution a été réalisée par 5 passages successifs de 0,5 volume de tampon E (composé de 100 mM de Tris, 150 mM de NaCl et 2,5 mM de D-desthiobiotine, pH 8). Les cycles supplémentaires de purification ont été réalisés après régénération de la résine par 2 passages successifs avec 1,5 volume tampon R (tampon W additionné de 1 mM d'HABA) et 2 passages successifs avec 10 volumes de tampon W.

V.2.4.3.3. Chromatographie par exclusion stérique (SEC)

Après purification par IMAC, les protéines destinées à la détermination de structure par cristallographie aux rayons X ont été purifiées par chromatographie d'exclusion stérique sur une résine HiLoad 26/600 Superdex 75 pg (GE Healthcare), à un débit de 2,6 ml/min maintenu par une station BioLogic DuoFlow (BioRad). En bref, les fractions d'élution collectées après IMAC ont été rassemblées et chargées dans une boucle de 10 ml, puis injectées sur la résine en collectant des fractions de 1,5 ml. Les fractions attribuées au pic de protéines recombinantes homogènes et non oligomérisées ont été rassemblées.

V.2.4.4. Dénaturation-renaturation et purification des protéines insolubles

Les culots de lyse obtenus lors de l'expression des charpentes humaines insolubles après greffe des résidus fonctionnels des feuillets anti-GFP de C6 ou C8, ainsi que les variants de Sso7d greffés, ont été soumis à un protocole de dénaturation/renaturation en présence d'agents chaotropiques. Deux approches distinctes de renaturation ont été explorées après resuspension des fractions insolubles dans un tampon à pH 8 composé d'urée 8 M, de Tris 0,01 M et de NaCl 0,1 M, par élimination graduelle ou rapide de l'agent chaotrope.

V.2.4.4.1. Gradient de renaturation

Après resuspension en présence de 8 M d'urée, les protéines ressolubilisées ont été traitées sur une résine de Nickel HisTrap FF (GE Healthcare) connectée à une station BioLogic DuoFlow (BioRad). En bref, les protéines ont été chargées dans une boucle de 5 ml, injectées sur la résine, puis lavées par passage de 50 ml de tampon de resuspension à un débit de 5 ml/min. La renaturation des protéines a ensuite été autorisée par application d'un gradient à un débit de 1 ml/min, pour équilibrer progressivement la résine d'un tampon d'urée 6 M (urée 6 M, de Tris 0,02 M et de NaCl 0,5 M, à pH 7,4) à un tampon d'urée 1 M (urée 1 M, de Tris 0,02 M et de NaCl 0,5 M, à pH 7,4) pendant 1,5 heure. Finalement, l'élution des protéines capturées a été réalisée par collection des fractions lors du passage de tampon additionné d'imidazole à 250 mM. Les fractions attribuées au pic de protéines recombinantes ont été rassemblées après analyse des chromatogrammes.

Entre les passages de différentes protéines recombinantes, la colonne a été régénérée par application successive de 3 volumes d'eau ultrapure, 3 volumes de TBS additionné de 50 mM d'acide éthylène diamine tétraacétique (EDTA), 3 volumes d'eau, puis 3 volumes de NiSO₄ 100 mM. Une analyse réalisée *a posteriori* a indiqué que la régénération dans ces conditions ne permet pas d'éliminer efficacement l'ensemble des protéines fusionnées à une étiquette poly-histidine retenues sur la résine, présentant un risque de contamination entre les passages.

V.2.4.4.2. Dilution rapide

Après resuspension en présence de 8 M d'urée, 4 ml de protéines ressolubilisées ont été diluées au goutte à goutte dans 40 ml de tampon A, sous agitation. Après dilution du volume total d'échantillon, les solutions diluées ont été incubées pendant 1 heure sous agitation, avant d'être purifiées de la façon décrite dans la section "V.2.4.3.1. Protéines avec étiquette poly-histidine (IMAC)".

V.2.4.5. Traduction *in vitro*

Pour accéder à une caractérisation rapide des protéines insolubles par expression en *E. coli*, sans nécessiter de lyse et purification, des micro-productions en solution moins complexe ont été réalisées à partir d'extrait de S30 couramment utilisé en *ribosome display* pour la traduction *in*

in vitro (Amstutz *et al.*, 2006). Les gènes codants pour les protéines à exprimer *in vitro* (Nanofitine H4 anti-lysozyme, Nanofitines anti-GFP C6 et C8, ainsi que les charpentes alternatives greffées des feuillettes de C6 et C8, en pAFG01) ont été amplifiés entre les séquences du promoteur T7 (insérée par l'oligonucléotide PURE_For) et du terminateur de transcription, en utilisant les mélanges réactionnels et cycles suivants (à l'aide d'un thermocycleur T100, BioRad):

<u>Composition du mélange réactionnel, en μl:</u>		<u>Programme PCR "INSERT5":</u>	
Eau ultrapure	36	98°C	30 sec
Phusion 5X buffer HF	10	98°C	10 sec
dNTP mix 10 mM	1	50°C	30 sec
Amorce PURE_For 10 μ M	1	72°C	45 sec
Amorce PURE_Rev 10 μ M	1	72°C	5 min
Phusion DNA polymerase	0,5	14°C	∞
Matrice (~250 ng/ μ l)	0,5		
Volume final	50		

} 30X

Un témoin négatif a été réalisé pour chaque réaction de PCR, dans les mêmes conditions en remplaçant la matrice par de l'eau ultrapure. Après analyse par électrophorèse sur gel d'agarose à 1,5% pendant 45 minutes à 110 V, les amplicons ont été purifiés à l'aide du kit de dessalage "Wizard® SV Gel and PCR Clean-Up System" (Promega) et élués dans 30 μ l d'eau ultrapure. Après dosage, les amplicons ont été soumis à une transcription *in vitro* à l'aide du kit "Kit TranscriptAid T7 High Yield Transcription" (Thermo Fisher), par incubation pendant 2 heures à 37°C dans le mélange réactionnel suivant:

Composition du mélange réactionnel, en μ l:

Eau ultrapure	Qsp 20
Tampon TranscriptAid 5X	4
NTP	8
Matrice ADN	1 μ g
Mélange enzymatique TranscriptAid	2
Volume final	20

Après transcription, les mélanges réactionnels ont été complétés avec 2 μ l de DNaseI et incubés pendant 15 minutes à 37°C, puis complétés avec 2 μ l d'EDTA 0,5 M à pH 8 et incubés pendant 10 minutes à 65°C. L'ARN transcrit a été purifié à l'aide du kit de dessalage "RNeasy MinElute Cleanup" (Qiagen) dans 16 μ l d'eau ultrapure, puis traduit après dosage par incubation pendant 1 heure à 37°C dans le mélange réactionnel suivant:

Composition du mélange réactionnel, en μ l:

Eau ultrapure	Qsp 82,5
ARN	7,5 μ g
Méthionine 200 mM	1,5
Pré-mix de traduction sans méthionine	30,75
Extrait de S30	37,5
Volume final	82,5

Finalement, chaque réaction a été arrêtée par ajout de 527,5 µl de TBS additionné de Tween 20 à 0,1%, aboutissant à la solution de protéines recombinantes traduites *in vitro*.

V.2.5. Caractérisations des protéines exprimées

V.2.5.1. SDS-PAGE

Après dosage à partir des valeurs d'absorbance à 280 nm déterminées par mesure entre 220 et 350 nm à l'aide d'un spectrophotomètre Varioskan (Thermo Scientific), 2 µg de protéines recombinantes purifiées ont été dilués dans 16 µl d'eau ultrapure, puis chauffés pendant 5 minutes à 95°C après addition de 4 µl de tampon de chargement non réducteur 5X. Les échantillons ainsi dénaturés ont été déposés sur gel de polyacrylamide à 15%, puis migrés par électrophorèse pendant 65 minutes à 150V en tampon Tris-glycine SDS. Les gels ont été ensuite colorés au bleu de Coomassie puis décolorés jusqu'à révélation de bandes protéiques distinctes. Les fractions non retenues de purification par IMAC et les fractions insolubles après lyse ont également été soumises à ce protocole en déposant, respectivement, 16 µl de solution ou un extrait collecté à la pointe d'un cône.

V.2.5.2. ELISA

V.2.5.2.1. Validation des clones anti-GFP purifiés issus de sélection *in vitro*

La fixation entre la GFP et les Nanofitines exprimées en erlenmeyer puis purifiées par IMAC a été validée par ELISA en présence ou en absence de la StrepTagII-GFP biotinylée utilisée lors de la sélection par *ribosome display*. En bref, la streptavidine (100 µl à 66 nM, Sigma-Aldrich) en TBS a été immobilisée dans les puits d'une plaque adhérente Maxisorp (Nunc) par incubation à 4°C pendant la nuit. Toutes les étapes suivantes ont été réalisées à température ambiante, avec une agitation à 600 rpm pour les étapes d'incubation. Les puits ont été lavés 3 fois avec 300 µl de TBS, puis bloqués avec 300 µl d'albumine bovine sérique (BSA, Sigma-Aldrich) en TBS pendant 1 heure. Le contenu des puits a été éliminé par retournement rapide et l'immobilisation de la StrepTagII-GFP biotinylée (100 µl à 40 nM) a été réalisée par incubation pendant 1 heure en TBS additionné de BSA 0,5%. Avant chacune des étapes d'incubation suivantes, les puits ont été lavés 3 fois avec 300 µl de TBS contenant 0,1% de Tween 20. Les Nanofitines purifiées (100 µl, dilués à 250 nM en TBS contenant 0,1% de Tween 20) ont été mises dans des puits avec ou sans antigène immobilisé pendant 1 heure.

La révélation a été effectuée par addition de 100 µl d'anticorps anti-RGS His conjugué à la HRP (Qiagen) dilué au 1:4000 en TBS contenant 0,1% de Tween 20 pendant 1 heure, suivie de l'ajout de 100 µl d'une solution de substrat *o*-Phenylenediamine dihydrochloride (OPD, Sigma-Aldrich) à 1 mg/ml en tampon de révélation (0,05 M acide citrique, 0,05% peroxyde d'hydrogène). L'absorbance à 450 nm a été mesurée à l'aide d'un lecteur de plaque ELISA Varioskan (Thermo Scientific). Cette méthode de révélation a été appliquée à l'ensemble des conditions ELISA décrites

ci-après, réalisées avec physisorption directe de StrepTagII-GFP non biotinylée (ou de ses variants) à 0,34 μM en remplacement de la streptavidine.

V.2.5.2.2. Variants alanine de la Nanofitine anti-GFP C8

La cartographie par mutations en alanine de la Nanofitine anti-GFP C8 a été réalisée par ELISA de la manière décrite dans la section “V.2.5.2.1. Validation des clones anti-GFP purifiés issus de sélection *in vitro*”. Les variants purifiés par IMAC ont été utilisés à des concentrations finales de Nanofitine de 4000; 800; 160; 32; 6,4; 1,28; 0,256; et 0,0512 nM afin de déterminer l'impact des mutations par substitution sur la capacité de fixation à la GFP.

V.2.5.2.3. Nanofitines conçues *de novo* (NF1 à NF10)

Les Nanofitines conçues *de novo* ont été criblées et caractérisées par ELISA sous trois formats différents: monomérique, tétramérique linéaire (fusion génétique) et tétramérique par couplage biotine/streptavidine. Les ELISA ont été réalisés de la manière décrite dans la section “V.2.5.2.1. Validation des clones anti-GFP purifiés issus de sélection *in vitro*”.

Les Nanofitines purifiées par IMAC ont été utilisées pour les deux premiers modes de présentation, à des concentrations finales de 1 et 10 μM pour les monomères, ou de 0,2 à 17 μM pour les multimères linéaires. Les multimères biotinylés ont été obtenus par biotinylation des Nanofitines, selon la méthode décrite dans la section “V.2.2.1. Biotinylation des antigènes”, aboutissant à deux molécules de biotine en moyenne par molécule de protéine et des solutions de Nanofitines de 150 à 800 μM . Les Nanofitines biotinylées (diluées à 10 μM) ont ensuite été incubées en TBS additionné de Tween 20 à 0,01% en présence de streptavidine (2,5 μM) pendant 1 heure à température ambiante sous agitation à 600 rpm. Les Nanofitines complexées ont ensuite été utilisées en ELISA à 1 μM ou 10 μM , équivalant à des concentrations de tétramères pouvant aller, respectivement jusqu'à 0,25 ou 2,5 μM .

Lors de la validation de l'épitope ciblé par NF5 en ELISA, la concentration du multimère linéaire de NF5 (NF5x4) a été fixée à 20 μM en présence des différents variants alanine de la StrepTagII-GFP.

V.2.5.3. Interférométrie de couche biologique

V.2.5.3.1. Détermination des affinités des Nanofitines

Les paramètres cinétiques de liaison entre les Nanofitines anti-GFP purifiées et la StrepTagII-GFP ont été mesurés par interférométrie sur un système Octet RED96 (ForteBio). Les Nanofitines, fusionnées à une étiquette poly-histidine et diluées à 25 $\mu\text{g/ml}$ en TBS, ont été chargées pendant 5 minutes à la surface de biocapteurs Ni-NTA équilibrés en TBS avec un décalage du motif de longueur d'onde compris entre 2 et 2,5 nm. Les biocapteurs, ensuite équilibrés pendant 3 minutes, ont alors été simultanément exposés à différentes concentrations de StrepTagII-GFP (100, 50, 25, 12,5, 6,25, 3,125, 1,563 et 0 nM). Les étapes d'association et de dissociation ont été mesurées

pendant 15 minutes chacune. Sauf précision contraire, toutes les étapes ont été réalisées en tampon TBS additionné de 0,002% Tween 20 et 0,01% BSA (TBS-T-BSA). Les biocapteurs ont été régénérés par 3 cycles d'incubations successives de 5 secondes en glycine 10 mM à pH 2 puis en TBS. Toutes les étapes ont été effectuées à 30°C avec une agitation continue à 1000 rpm. Le biocapteur exposé à une concentration de 0 nM a été utilisé comme bruit de fond de référence. Les sensorgrammes ont été analysés à l'aide du logiciel Octet Data Analysis version 7.1 (ForteBio), en soustrayant le signal de référence et en utilisant un modèle d'ajustement 1:1 pour les étapes d'association et dissociation.

V.2.5.3.2. Détermination d'épitopes chevauchants par compétition en tandem

Le regroupement des Nanofitines anti-GFP par épitope chevauchant à la surface de la StrepTagII-GFP a été réalisé par compétition en tandem et mesuré par interférométrie sur un système Octet RED96 (ForteBio). La StrepTagII-GFP, biotinylée et diluée à 5 µg/ml en TBS, a été chargée pendant 10 minutes à la surface de biocapteurs Streptavidine équilibrés en TBS avec un décalage du motif de longueur d'onde compris entre 1,6 et 2,0 nm. Les biocapteurs, ensuite équilibrés pendant 3 minutes, ont alors été simultanément saturés par incubation pendant 15 minutes avec une première Nanofitine à 300 nM (Nanofitine saturante), avant incubation pendant 10 minutes avec une seconde Nanofitine à 100 nM (Nanofitine compétitrice). Une étape finale de dissociation a également été mesurée pendant 3 minutes. Sauf précision contraire, toutes les étapes ont été réalisées en tampon TBS additionné de 0,002% Tween 20 et 0,01% BSA (TBS-T-BSA). Les biocapteurs ont été régénérés par 3 cycles d'incubations successives de 5 secondes en HCl à pH 3 puis en TBS. Toutes les étapes ont été effectuées à 30°C avec une agitation continue à 1000 rpm. Lors de chaque cycle, un biocapteur a été exposé à une concentration de 0 nM de Nanofitine saturante pour confirmer la capacité de fixation de la Nanofitine compétitrice. Les sensorgrammes ont été analysés à l'aide du logiciel Octet Data Analysis version 7.1 (ForteBio), en utilisant un modèle d'ajustement 1:1 pour les étapes d'association et dissociation.

V.2.5.3.3. Fixation des protéines après transfert de feuillet anti-GFP

La fixation à la StrepTagII-GFP par les Nanofitines anti-GFP C6 et C8, les charpentes alternatives greffées de leurs résidus fonctionnels et les constructions chimériques fusionnées avec les protéines TrxA, NusA ou MBP a été caractérisée par interférométrie sur un système Octet RED96 (ForteBio). En bref, le protocole appliqué correspond à celui décrit dans la section "V.2.5.3.1. Détermination des affinités des Nanofitines", réalisé avec 0,5 ou 5 µM de StrepTagII-GFP et des étapes d'association et de dissociation de 5 minutes chacune.

V.2.5.3.4. Cartographie de l'interaction C8:GFP par mutations en alanine

Les paramètres cinétiques de liaison entre les variants alanine de la Nanofitine anti-GFP C8, purifiés par IMAC, et la StrepTagII-GFP ont été mesurés par interférométrie sur un système Octet

RED96 (ForteBio). Pour stabiliser les signaux obtenus au cours du temps, des biocapteurs ont été fonctionnalisés en immobilisant la Nanofitine anti-GFP C6 (fixant un épitope différent de celui de C8) avec une orientation contrôlée par capture à la surface de biocapteurs Ni-NTA via leur étiquette poly-histidine, puis en réalisant un couplage covalent. Toutes les étapes suivantes ont été effectuées à 30°C avec une agitation continue à 1000 rpm.

La Nanofitine C6, purifiée et diluée à 25 µg/ml en TBS, a été chargée pendant 30 minutes à la surface de biocapteurs Ni-NTA équilibrés en TBS avec un décalage du motif de longueur d'onde de 3 nm. Les biocapteurs ont ensuite été équilibrés pendant 30 secondes en PBS, puis activés par incubation pendant 10 minutes avec une solution de EDC:Sulfo-NHS (100 µl de chlorhydrate de 1-ethyl-3-(3-diméthylaminopropyl)carbodiimide à 0,4 M, Fisher Scientific; 100 µl de N-Hydroxysulfosuccinimide à 0,2 M, Fisher Scientific). Le blocage des biocapteurs a été réalisé par incubation pendant 10 minutes avec de l'éthanolamine (Sigma-Aldrich) diluée à 0,5 M en solution de NaCl 0,5 M à pH 8,3, puis le nickel a été éliminé par chélation pendant 10 minutes en présence d'acide éthylènediaminetétraacétique à 0,5 M (EDTA, Merck Millipore).

La StrepTagII-GFP, diluée à 25 µg/ml, a été chargée pendant 5 minutes à la surface de biocapteurs fonctionnalisés avec C6 et équilibrés avec un décalage du motif de longueur d'onde compris entre 0,5 et 0,7 nm. Les biocapteurs, ensuite équilibrés pendant 3 minutes, ont alors été simultanément exposés à différentes concentrations de variant de la Nanofitine C8 (100, 50, 25, 12,5, 6,25, 3,125, 1,563 et 0 nM). Les étapes d'association et de dissociation ont été mesurées pendant 10 minutes chacune. Sauf précision contraire, toutes les étapes ont été réalisées en tampon TBS additionné de 0,002% Tween 20 et 0,01% BSA (TBS-T-BSA). Les biocapteurs ont été régénérés par 5 cycles d'incubations successives de 10 secondes en glycine 10 mM à pH 2 puis en TBS. Toutes les étapes ont été effectuées à 30°C avec une agitation continue à 1000 rpm. Le biocapteur exposé à une concentration de 0 nM a été utilisé comme bruit de fond de référence. Les sensorgrammes ont été analysés à l'aide du logiciel Octet Data Analysis version 7.1 (ForteBio), en soustrayant le signal de référence et en utilisant un modèle d'ajustement 1:1 pour les étapes d'association et de dissociation. Malgré des résultats d'association en corrélation avec les signaux observés en ELISA, les constantes d'affinités exactes n'ont pas été déterminées en raison de mauvais ajustements sur l'étape de dissociation.

V.2.5.4. Mesures de fluorescence

V.2.5.4.1. Décalage de fluorescence induit par la fixation de Nanofitine

Les spectres d'émission de fluorescence de la StrepTagII-GFP en absence ou présence de Nanofitines anti-GFP ont été déterminés à partir des huit Nanofitines générées *in vitro*. La StrepTagII-GFP purifiée par chromatographie d'affinité (2,5 µM en tampon E) a été déposée en microplaques 96-puits noires non-adhérentes (Greiner Bio One), en présence de Nanofitines purifiées par IMAC (40 µM en tampon B) dans un volume final de 100 µl de TBS. L'émission de

fluorescence entre 495 et 540 nm a été mesurée à l'aide d'un lecteur de plaque Infinite M1000 (Tecan), après excitation à 475 nm. La comparaison des différentes conditions a été réalisée sur la base des longueurs d'onde maximales d'émission de fluorescence à partir de 50 mesures par échantillon, sans inclure l'intensité de fluorescence observée pour s'affranchir du phénomène de photoblanchiment.

V.2.5.4.2. Spectres d'émission de fluorescence des variants GFP

Les spectres d'émission de fluorescence de la StrepTagII-GFP et ses variants alanine ont été déterminés pour confirmer leur quantification et leur bon repliement, notamment pour assurer la normalisation des ELISA réalisés avec NF5 tétramérique (NF5x4). Les variants de StrepTagII-GFP purifiés par chromatographie d'affinité (200 µl dilués à 1 µM en TBS) ont été déposés en microplaques 96-puits noires non-adhérentes (Greiner Bio One). L'émission de fluorescence entre 490 et 520 nm a été mesurée à l'aide d'un lecteur de plaque Varioskan (Thermo Scientific), après excitation à 475 nm. Les variants de la GFP ont été considérés non affectés par l'introduction de mutation en cas de spectre présentant un pic de maximum d'émission de fluorescence entre 500 et 505 nm d'environ 700 unités relative de fluorescence (rfu).

V.2.5.5. Criblage des conditions de cristallogenèse

Trois conditions de cristallogenèse ont été explorées avec le soutien de l'équipe de Leonardo Scapozza (Université de Genève, Suisse), via la participation d'Andreja Vujicic-Zagar et de Magali Zeisser-Labouèbe en particulier. Ainsi, 384 solutions de précipitants commerciaux ont été criblés par la méthode de la goutte assise (ou *sitting drop*) avec la Nanofitine NF5 sous forme monomérique ainsi que les chimères NF5-L7-GFP et NF5-L20-GFP. En bref, 35 µl de chaque solution commerciale de composition variable en précipitant ont été déposés par puits, et des gouttes de protéine ont été déposées sur les trois sellettes de chaque puits à l'aide d'un distributeur automatique Nanodrop Express Innovadyne (Solve Scientific), après dilution à un ratio protéine:précipitant de 200:200, 275:125 et 125:275 (en µl). L'observation de l'évolution individuelle des gouttes déposées a été réalisée pendant 5 semaines pour identifier des conditions favorables à la formation de cristaux, à l'aide du logiciel Xtal Focus pilotant l'automate de capture d'image (ExploraNova). Les précipitants criblés ont été préparés à partir des kits "JCSG-Plus" (Molecular Dimensions), "Modern Intelligent Dynamic Alternative Screen (MIDAS)" (Molecular Dimensions), "Crystal Screen HT" (Hampton Research) et "PEG/Ion HT" (Hampton Research).

Bibliographie

Bibliographie

- Abuchowski, a., McCoy, J. R., Palczuk, N. C., van Es, T., & Davis, F. F. (1977). Effect of covalent attachment of polyethylene glycol on immunogenicity and circulating life of bovine liver catalase. *Journal of Biological Chemistry*, 252(11), 3582-3586.
- Agilent-Technologies. (1991). U.S. Patent 6,713,285; 6,391,548; 5,948,663; 5,932,419; 5,789,166; 7,132,265; 7,176,004; 5,286,632; and patents pending.
- Aldington, S., & Bonnerjea, J. (2007). Scale-up of monoclonal antibody purification processes. *Journal of Chromatography, B: Analytical Technologies in the Biomedical and Life Sciences*, 848(1), 64-78.
- Amstutz, P., Binz, H. K., Zahnd, C., & Plückchun, A. (2006). Ribosome Display: In Vitro Selection of Protein-Protein Interactions. In J. E. Celis (Éd.), *Cell Biology, A Laboratory Handbook* (3^e éd., Vol. 1, p. 497-503). Elsevier Academic Press.
- AntibodyRegistry. (2015). List of registred Anti-GFP Antibodies. Consulté à l'adresse <http://antibodyregistry.org/search?q=anti-GFP> (consulté le 16 août 2015)
- Asherie, N. (2004). Protein crystallization and phase diagrams. *Methods*, 34(3), 266-272.
- Azoitei, M. L., Correia, B. E., Ban, Y.-E. a., Carrico, C., Kalyuzhniy, O., Chen, L., ... Schief, W. R. (2011). Computation-Guided Backbone Grafting of a Discontinuous Motif onto a Protein Scaffold - Supplementary Online Material. *Science*, 334(6054), 373-376.
- Babor, M., Mandell, D. J., & Kortemme, T. (2011). Assessment of flexible backbone protein design methods for sequence library prediction in the therapeutic antibody Herceptin-HER2 interface. *Protein Science*, 20(6), 1082-1089.
- Baker, M., Reynolds, H. M., Lusicisi, B., & Bryson, C. J. (2010). Immunogenicity of protein therapeutics: The key causes, consequences and challenges. *Self/Nonself*, 1(4), 314-322.
- Balint, R. F., & Larrick, J. W. (1993). Antibody engineering by parsimonious mutagenesis. *Gene*, 137(1), 121-126.
- Barderas, R., Desmet, J., Timmerman, P., Meloen, R., & Casal, J. I. (2008). Affinity maturation of antibodies assisted by in silico modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 105(26), 9029-9034.
- Baskaran, K., Duarte, J. M., Biyani, N., Bliven, S., & Capitani, G. (2014). A PDB-wide , evolution-based assessment of protein – protein interfaces. *BMC Structural Biology*, 12, 22.
- Beck, A., Wurch, T., Bailly, C., & Corvaia, N. (2010). Strategies and challenges for the next generation of therapeutic antibodies. *Nature reviews. Immunology*, 10(5), 345-52.
- Béhar, G., Bellinzoni, M., Maillasson, M., Paillard-Laurance, L., Alzari, P. M., He, X., ... Pecorari, F. (2013). Tolerance of the archaeal Sac7d scaffold protein to alternative library designs: characterization of anti-immunoglobulin G Affitins. *Protein engineering, design & selection : PEDS*, 1-9.

- Béhar, G., Pacheco, S., Maillason, M., Mouratou, B., & Pecorari, F. (2014). Switching an anti-IgG binding site between archaeal extremophilic proteins results in Affitins with enhanced pH stability. *Journal of biotechnology*, 192 Pt A, 123-9.
- Benedix, A., Becker, C. M., de Groot, B. L., Caflisch, A., & Böckmann, R. A. (2009). Predicting free energy changes using structural ensembles. *Nature methods*.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1), 235-242.
- Binz, H. K., Amstutz, P., & Plückthun, A. (2005). Engineering novel binding proteins from nonimmunoglobulin domains. *Nature biotechnology*, 23(10), 1257-68.
- Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15), 5869-5874.
- Bocci, V. (1990). Catabolism of therapeutic proteins and peptides with implications for drug delivery. *Advanced drug delivery reviews*, 4, 149-169.
- Boersma, Y. L., & Plückthun, A. (2011). DARPins and other repeat protein scaffolds: advances in engineering and applications. *Current opinion in biotechnology*, 22(6), 849-57.
- Bostrom, J., Yu, S.-F., Kan, D., Appleton, B. a, Lee, C. V, Billeci, K., ... Fuh, G. (2009). Variants of the antibody herceptin that interact with HER2 and VEGF at the antigen binding site. *Science (New York, N.Y.)*, 323(5921), 1610-1614.
- Brauchle, M., Hansen, S., Caussin, E., Lenard, A., Ochoa-Espinosa, A., Scholz, O., ... Affolter, M. (2014). Protein interference applications in cellular and developmental biology using DARPins that recognize GFP and mCherry. *Biology open*, 3(12), 1252-61.
- Buddelmeijer, N., Krehenbrink, M., Pecorari, F., & Pugsley, A. P. (2009). Type II secretion system secretin PulD localizes in clusters in the Escherichia coli outer membrane. *Journal of bacteriology*, 191(1), 161-8.
- Caldas, C., Coelho, V., Kalil, J., Moro, A. M., Maranhão, A. Q., & Brígido, M. M. (2003). Humanization of the anti-CD18 antibody 6.7: An unexpected effect of a framework residue in binding to antigen. *Molecular Immunology*, 39(15), 941-952.
- Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1), 41-51.
- Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H., & Gray, J. J. (2011). Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PloS one*, 6(8), e22477.
- Chen, J., Sawyer, N., & Regan, L. (2013). Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Science*, 22(4), 510-515.
- Chien, C. T., Bartel, P. L., Sternglanz, R., & Fields, S. (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proceedings of the National Academy of Sciences of the United States of America*, 88(21), 9578-82.

- Chin, J. W., & Schepartz, A. (2001). Design and evolution of a miniature Bcl-2 binding protein. *Angewandte Chemie - International Edition*, 40(20), 3806-3809.
- Choi, Y., Griswold, K. E., & Bailey-Kellogg, C. (2013). Structure-based redesign of proteins for minimal T-cell epitope content. *Journal of computational chemistry*.
- Chothia, C., & Janin, J. (1975). Principles of protein-protein recognition. *Nature*, 256(5520), 705-8.
- Cinier, M., Labouebe, M., Rousseaux, C., Rodriguez, J., Cunha, A., Tan, N., ... Kitten, O. (2015). Oral delivery of a new class of non-antibody protein scaffold Nanofitins targeting TNF-alpha shows a strong preventive and curative anti-inflammatory effect in models of inflammatory bowel diseases. *Journal of Crohn's and Colitis*, 9(suppl 1), S17-S18.
- Cinier, M., Petit, M., Pecorari, F., Talham, D. R., Bujoli, B., & Tellier, C. (2012). Engineering of a phosphorylatable tag for specific protein binding on zirconium phosphonate based microarrays. *JBIC, Journal of Biological Inorganic Chemistry*, 17(3), 399-407.
- Cinier, M., Petit, M., Williams, M. N., Fabre, R. M., Pecorari, F., Talham, D. R., ... Tellier, C. (2009). Bisphosphonate adaptors for specific protein binding on zirconium phosphonate-based microarrays. *Bioconjugate chemistry*, 20(12), 2270-7.
- Clark, L. a, Boriack-Sjodin, P. A., Eldredge, J., Fitch, C., Friedman, B., Hanf, K. J. M., ... Van Vlijmen, H. (2006). Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein science : a publication of the Protein Society*, 15(5), 949-960.
- Combs, S. a, Deluca, S. L., Deluca, S. H., Lemmon, G. H., Nannemann, D. P., Nguyen, E. D., ... Meiler, J. (2013). Small-molecule ligand docking into comparative models with Rosetta. *Nature protocols*, 8(7), 1277-98.
- Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*.
- Correa, A., Pacheco, S., Mechaly, A. E., Obal, G., Béhar, G., Mouratou, B., ... Pecorari, F. (2014). Potent and Specific Inhibition of Glycosidases by Small Artificial Binding Proteins (Affitins). *PLoS ONE*, 9(5), e97438.
- Correia, B. E., Ban, Y. E. A., Holmes, M. a., Xu, H., Ellingson, K., Kraft, Z., ... Schief, W. R. (2010). Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. *Structure*, 18(9), 1116-1126.
- Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., ... Schief, W. R. (2014). Proof of principle for epitope-focused vaccine design. *Nature*, 507(7491), 201-6.
- Corti, D., Voss, J., Gamblin, S. J., Codoni, G., Macagno, A., Jarrossay, D., ... Lanzavecchia, A. (2011). A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science (New York, N.Y.)*, 333(6044), 850-856.
- Cossio, P., Trovato, A., Pietrucci, F., Seno, F., Maritan, A., & Laio, A. (2010). Exploring the universe of protein structures beyond the protein data bank. *PLoS Computational Biology*, 6(11).
- Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Research*, 14(6), 1188-1190.

- Dahiyat, B. I., & Mayo, S. L. (1997). De novo protein design: towards fully automated sequence selection. *Journal of molecular biology*, 278(5335), 82-87.
- Daily, M. D., Masica, D., Sivasubramanian, A., Somarouthu, S., & Gray, J. J. (2005). CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock. *Proteins: Structure, Function and Genetics*, 60(2), 181-186.
- Damborsky, J., & Brezovsky, J. (2014). Computational tools for designing and engineering enzymes. *Current Opinion in Chemical Biology*, 19(1), 8-16.
- Dantas, G., Kuhlman, B., Callender, D., Wong, M., & Baker, D. (2003). A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *Journal of Molecular Biology*, 332(2), 449-460.
- Das, R., André, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., ... Baker, D. (2009). Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), 18978-18983.
- Das, R., & Baker, D. (2008). Macromolecular modeling with rosetta. *Annual review of biochemistry*, 77, 363-382.
- Daura, X., Gademann, K., Jaun, B., Seebach, D., Van Gunsteren, W. F., & Mark, A. E. (1999). Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition*, 38(1-2), 236-240.
- Davis, G. D., Elisee, C., Mewham, D. M., & Harrison, R. G. (1999). New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnology and Bioengineering*, 65(4), 382-388.
- De Beer, T. a P., Berka, K., Thornton, J. M., & Laskowski, R. a. (2014). PDBsum additions. *Nucleic Acids Research*, 42(D1), 292-296.
- De Marco, A. (2009). Strategies for successful recombinant expression of disulfide bond-dependent proteins in Escherichia coli. *Microbial cell factories*, 8, 26.
- Deehan, M., Garcês, S., Kramer, D., Baker, M. P., Rat, D., Roettger, Y., & Kromminga, A. (2015). Managing unwanted immunogenicity of biologicals. *Autoimmunity Reviews*, 14(7), 569-574.
- DeLano, W. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsletter On Protein Crystallography*.
- Denisova, G., Lideman, L., Spectorman, E., Abulafia-Lapid, R., Burke, M., Yust, I., & Gershoni, J. M. (2003). Characterization of new monoclonal antibodies that discriminate between soluble and membrane CD4 and compete with human anti-CD4 autoimmune sera. *Molecular Immunology*, 40(5), 231-239.
- Dennis, M. S., Zhang, M., Gloria Meng, Y., Kadkhodayan, M., Kirchhofer, D., Combs, D., & Damico, L. a. (2002). Albumin binding as a general strategy for improving the pharmacokinetics of proteins. *Journal of Biological Chemistry*, 277(38), 35035-35043.
- Der, B. S., Kluwe, C., Miklos, A. E., Jacak, R., Lyskov, S., Gray, J. J., ... Kuhlman, B. (2013). Alternative Computational Protocols for Supercharging Protein Surfaces for Reversible Unfolding and Retention of Stability. *PLoS ONE*, 8(5).

- Desjarlais, J., Zalevsky, J., & Moore, G. (2004). US Patent App. 10/956,352: Methods for rational pegylation of proteins.
- DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W., & Baker, D. (2009). Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta. *Journal of Molecular Biology*, 392(1), 181-190.
- Dimitrov, D. S. (2012). Therapeutic proteins. *Methods in molecular biology (Clifton, N.J.)*, 899, 1-26.
- Dourado, D. F. a R., & Flores, S. C. (2014). A multiscale approach to predicting affinity changes in protein-protein interfaces. *Proteins*, (March), 1-10.
- Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R., & Meiler, J. (2009). Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *Journal of Molecular Modeling*, 15(9), 1093-1108.
- Edmondson, S. P., & Shriver, J. W. (2001). DNA binding proteins Sac7d and Sso7d from *Sulfolobus*. *Methods in enzymology*, 334(1986), 129-45.
- Ekiert, D. C., Bhabha, G., Elsliger, M., Friesen, R. H. E., Jongeneelen, M., Throsby, M., ... Wilson, I. a. (2009). Antibody recognition of a highly conserved influenza virus epitope. *Science (New York, N.Y.)*, 324(5924), 246-251.
- Ellis, L. M., & Hicklin, D. J. (2008). VEGF-targeted therapy: mechanisms of anti-tumour activity. *Nature reviews. Cancer*, 8(8), 579-91.
- Even-Desrumeaux, K., Nevoltris, D., Lavaut, M. N., Alim, K., Borg, J.-P., Audebert, S., ... Chames, P. (2014). Masked selection: a straightforward and flexible approach for the selection of binders against specific epitopes and differentially expressed proteins by phage display. *Molecular & cellular proteomics : MCP*, 13(2), 653-65.
- Farady, C. J., Sellers, B. D., Jacobson, M. P., & Craik, C. S. (2009). Improving the species cross-reactivity of an antibody using computational design. *Bioorganic and Medicinal Chemistry Letters*, 19(14), 3744-3747.
- Fazelinia, H., Cirino, P. C., & Maranas, C. D. (2009). OptGraft: A computational procedure for transferring a binding site onto an existing protein scaffold. *Protein Science*, 18(1), 180-195.
- Fiedler, M., & Skerra, A. (2008). *Non-Antibody Scaffolds. Handbook of Therapeutic Antibodies*. (S. Dübel, Éd.). Weinheim, Germany: Wiley-VCH Verlag GmbH.
- Fields, S., & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230), 245-6.
- Filmore, D. (2004). It's a GPCR world. *Modern drug discovery*, 7(11), 24-27.
- Fleishman, S. J., Corn, J. E., Strauch, E. M., Whitehead, T. a., Andre, I., Thompson, J., ... Baker, D. (2010). Rosetta in CAPRI rounds 13-19. *Proteins: Structure, Function and Bioinformatics*, 78(15), 3212-3218.
- Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E., ... Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science (New York, N.Y.)*, 332(6031), 816-821.

- Fogolari, F., Brigo, a., & Molinari, H. (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: A tool for structural biology. *Journal of Molecular Recognition*, 15(6), 377-392.
- Folli, C., Calderone, V., Ramazzina, I., Zanotti, G., & Berni, R. (2002). Ligand binding and structural analysis of a human putative cellular retinol-binding protein. *The Journal of biological chemistry*, 277(44), 41970-7.
- Friesen, R. H. E., Lee, P. S., Stoop, E. J. M., Hoffman, R. M. B., Ekiert, D. C., Bhabha, G., ... Wilson, I. a. (2014). A common solution to group 2 influenza virus neutralization. *Proceedings of the National Academy of Sciences of the United States of America*, 111(1), 445-50.
- Gainza, P., Roberts, K. E., Georgiev, I., Lilien, R. H., Keedy, D. A., Chen, C. Y., ... Donald, B. R. (2013). Osprey: Protein design with ensembles, flexibility, and provable algorithms. *Methods in Enzymology*, 523(7405), 87-107.
- Garcia-Rodriguez, C., Levy, R., Arndt, J. W., Forsyth, C. M., Razai, A., Lou, J., ... Marks, J. D. (2007). Molecular evolution of antibody cross-reactivity for two subtypes of type A botulinum neurotoxin. *Nature biotechnology*, 25(1), 107-116.
- Gebauer, M., & Skerra, A. (2009). Engineered protein scaffolds as next-generation antibody therapeutics. *Current opinion in chemical biology*, 13(3), 245-55.
- Gera, N., Hill, A. B., White, D. P., Carbonell, R. G., & Rao, B. M. (2012). Design of pH sensitive binding proteins from the hyperthermophilic Sso7d scaffold. *PLoS one*, 7(11), e48928.
- Gera, N., Hussain, M., Wright, R. C., & Rao, B. M. (2011). Highly stable binding proteins derived from the hyperthermophilic Sso7d scaffold. *Journal of molecular biology*, 409(4), 601-16.
- Gerstein, M. (1998). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding and Design*, 3(6), 497-512.
- Gibson, D. G., Smith, H. O., Hutchison, C. a, Venter, J. C., & Merryman, C. (2010). Chemical synthesis of the mouse mitochondrial genome. *Nature methods*, 7(11), 901-3.
- Gillessen, S., Carvajal, D., Ling, P., Podlaski, F. J., Stremlo, D. L., Familletti, P. C., ... Gately, M. K. (1995). Mouse interleukin-12 (IL-12) p40 homodimer: A potent IL-12 antagonist. *European Journal of Immunology*, 25(1), 200-206.
- Goeddel, D. V., Kleid, D. G., Bolivar, F., Heyneker, H. L., Yansura, D. G., Crea, R., ... Riggs, a D. (1979). Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences of the United States of America*, 76(1), 106-110.
- Gottesman, S., Gottesman, M., Shaw, J. E., & Pearson, M. L. (1981). Protein degradation in E. coli: The lon mutation and bacteriophage lambda N and cII protein stability. *Cell*, 24(1), 225-233.
- Gray, J. J., Moughon, S. E., Kortemme, T., Schueler-Furman, O., Misura, K. M. S., Morozov, A. V, & Baker, D. (2003). Protein-protein docking predictions for the CAPRI experiment. *Proteins*, 52(1), 118-22.
- Griffin, L., & Lawson, a. (2011). Antibody fragments as tools in crystallography. *Clinical and Experimental Immunology*, 165(3), 285-291.

- Grodberg, J., & Dunn, J. J. (1988). ompT encodes the Escherichia coli outer membrane protease that cleaves T7 RNA polymerase during purification. *Journal of Bacteriology*, 170(3), 1245-1253.
- Grönwall, C., & Ståhl, S. (2009). Engineered affinity proteins - Generation and applications. *Journal of biotechnology*, 140(3-4), 254-69.
- Gubler, U., Chua, a O., Schoenhaut, D. S., Dwyer, C. M., McComas, W., Motyka, R., ... Familletti, P. C. (1991). Coexpression of two distinct genes is required to generate secreted bioactive cytotoxic lymphocyte maturation factor. *Proceedings of the National Academy of Sciences of the United States of America*, 88(10), 4143-4147.
- Guellouz, A., Valerio-Lepiniec, M., Urvoas, A., Chevrel, A., Graille, M., Fourati-Kammoun, Z., ... Minard, P. (2013). Selection of Specific Protein Binders for Pre-Defined Targets from an Optimized Library of Artificial Helicoidal Repeat Proteins (alphaRep). *PLoS ONE*, 8(8).
- Gunasekera, S., Foley, F. M., Clark, R. J., Sando, L., Fabri, L. J., Craik, D. J., & Daly, N. L. (2008). Engineering stabilized vascular endothelial growth factor-A antagonists: Synthesis, structural characterization, and bioactivity of grafted analogues of cyclotides. *Journal of Medicinal Chemistry*, 51(24), 7697-7704.
- Guntas, G., Purbeck, C., & Kuhlman, B. (2010). Engineering a protein-protein interface using a computationally designed library. *Proceedings of the National Academy of Sciences of the United States of America*, 107(45), 19296-19301.
- Handel, T., & DeGrado, W. (1990). De Novo Design of a Zn²⁺ binding protein. *Journal of the American Chemical Society*, 112, 6710-6711.
- Hanes, J., & Plückthun, A. (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), 4937-42.
- Hansen, F. G., Christensen, B. B., & Atlung, T. (1991). The initiator titration model: computer simulation of chromosome and minichromosome control. *Research in microbiology*, 142(2-3), 161-167.
- Heaney, M. L., & Golde, D. W. (1998). Soluble receptors in human disease Abstract : Soluble cytokine receptors naturally. *Journal of Leukocyte Biology*, 64(August), 135-146.
- Hecht, M. H., Richardson, J. S., Richardson, D. C., & Ogden, R. C. (1990). De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science (New York, N.Y.)*, 249(4971), 884-891.
- Henager, S. H., Hale, M. a., Maurice, N. J., Dunnington, E. C., Swanson, C. J., Peterson, M. J., ... McFarland, B. J. (2012). Combining different design strategies for rational affinity maturation of the MICA-NKG2D interface. *Protein Science*, 21(9), 1396-1402.
- Hey, T., Fiedler, E., Rudolph, R., & Fiedler, M. (2005). Artificial, non-antibody binding proteins for pharmaceutical and industrial applications. *Trends in biotechnology*, 23(10), 514-22.
- Holliger, P., & Hudson, P. J. (2005). Engineered antibody fragments and the rise of single domains. *Nature biotechnology*, 23(9), 1126-36.

- Holt, L. J., Basran, A., Jones, K., Chorlton, J., Jespers, L. S., Brewis, N. D., & Tomlinson, I. M. (2008). Anti-serum albumin domain antibodies for extending the half-lives of short lived drugs. *Protein Engineering, Design and Selection*, 21(5), 283-288.
- Hosse, R., Rothe, A., & Power, B. (2006). A new generation of protein display scaffolds for molecular recognition. *Protein science*, (Table 1), 14-27.
- Huet, S., Gorre, H., Perrocheau, A., Picot, J., & Cinier, M. (2015). Use of the Nanofitin alternative scaffold as a GFP-ready fusion tag. *PLoS ONE*, 10(11): e0142304.
- Humphris, E. L., & Kortemme, T. (2008). Prediction of Protein-Protein Interface Sequence Diversity Using Flexible Backbone Computational Protein Design. *Structure*, 16(12), 1777-1788.
- Hurle, M. R., & Gross, M. (1994). Protein engineering techniques for antibody humanization. *Current Opinion in Biotechnology*, 5(4), 428-433.
- Hussain, M., Gera, N., Hill, A. B., & Rao, B. M. (2012). Scaffold Diversification Enhances Effectiveness of a Superlibrary of Hyperthermophilic Proteins. *ACS Synthetic Biology*.
- Hussain, M., Lockney, D., Wang, R., Gera, N., & Rao, B. M. (2013). Avidity-mediated virus separation using a hyperthermophilic affinity ligand. *Biotechnology progress*, 29(1), 237-46.
- Inoue, H., Ihara, A., Takahashi, H., Shimada, I., Ishida, I., & Maeda, Y. (2011). Affinity transfer to a human protein by CDR3 grafting of camelid VHH. *Protein Science*, 20(12), 1971-1981.
- Inoue, H., Suganami, A., Ishida, I., Tamura, Y., & Maeda, Y. (2013). Affinity maturation of a CDR3-grafted VHH using in silico analysis and surface plasmon resonance. *Journal of Biochemistry*, 154(4), 325-332.
- Jacak, R., Leaver-Fay, A., & Kuhlman, B. (2012). Computational protein design with explicit consideration of surface hydrophobic patches. *Proteins: Structure, Function and Bioinformatics*, 80(3), 825-838.
- Jacquet, a, Daminet, V., Haumont, M., Garcia, L., Chaudoir, S., Bollen, a, & Biemans, R. (1999). Expression of a recombinant *Toxoplasma gondii* ROP2 fragment as a fusion protein in bacteria circumvents insolubility and proteolytic degradation. *Protein expression and purification*, 17(3), 392-400.
- Jain, P., Taxak, P., Dash, P., Gaikwad, K., Kansal, R., & Gupta, V. (2014). Next-Generation Sequencing: Principle and Applications to Crops. In *Omics Technologies and Crop Improvement* (p. 323-342). CRC Press.
- Janin, J. (1997). Specific versus non-specific contacts in protein crystals. *Nature structural biology*, 4(12), 973-974.
- Janin, J. (2002). Welcome to CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function and Genetics*, 47(3), 257.
- Jha, R. K., Leaver-Fay, A., Yin, S., Wu, Y., Butterfoss, G. L., Szyperski, T., ... Kuhlman, B. (2010). Computational Design of a PAK1 Binding Protein. *Journal of Molecular Biology*, 400(2), 257-270.

- Jiang, L., Althoff, E. a, Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., ... Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science (New York, N.Y.)*, 319(5868), 1387-91.
- Joachimiak, L. a., Kortemme, T., Stoddard, B. L., & Baker, D. (2006). Computational Design of a New Hydrogen Bond Network and at Least a 300-fold Specificity Switch at a Protein-Protein Interface. *Journal of Molecular Biology*, 361(1), 195-208.
- Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S., & Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*, 321(6069), 522-525.
- Kang, H. J., Lee, C., & Drew, D. (2013). Breaking the barriers in membrane protein crystallography. *The international journal of biochemistry & cell biology*, 45(3), 636-44.
- Kapoor, D., Kumar, V., Chandrayan, S. K., Ahmed, S., Sharma, S., Datt, M., ... Guptasarma, P. (2008). Replacement of the active surface of a thermophile protein by that of a homologous mesophile protein through structure-guided « protein surface grafting ». *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1784(11), 1771-1776.
- Kapoor, D., Singh, B., Subramanian, K., & Guptasarma, P. (2009). Creation of a new eye lens crystallin (Gambeta) through structure-guided mutagenic grafting of the surface of Beta-B2 crystallin onto the hydrophobic core of Gamma-B crystallin. *FEBS Journal*, 276(12), 3341-3353.
- Kapust, R. B., & Waugh, D. S. (1999). Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein science : a publication of the Protein Society*, 8(8), 1668-1674.
- Karanicolas, J., Corn, J. E., Chen, I., Joachimiak, L. A., Dym, O., Peck, S. H., ... Baker, D. (2011). A De Novo Protein Binding Pair By Computational Design and Directed Evolution. *Molecular Cell*, 42(2), 250-260.
- Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*.
- Keck, P. C., & Huston, J. S. (1996). Symmetry of Fv architecture is conducive to grafting a second antibody binding site in the Fv region. *Biophysical journal*, 71(4), 2002-2011.
- Khoury, G. a., Smadbeck, J., Kieslich, C. a., & Floudas, C. a. (2014). Protein folding and de novo protein design for biotechnological applications. *Trends in Biotechnology*, 32(2), 99-109.
- Kilambi, K. P., & Gray, J. J. (2012). Rapid calculation of protein pKa values using rosetta. *Biophysical Journal*, 103(3), 587-595.
- King, N. P., Sheffler, W., Sawaya, M. R., Vollmar, B. S., Sumida, J. P., Andre, I., ... Baker, D. (2012). Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science*, 336(6085), 1171-1174.
- Kirchhofer, A., Helma, J., Schmidthals, K., Frauer, C., Cui, S., Karcher, A., ... Rothbauer, U. (2010). Modulation of protein properties in living cells using nanobodies. *Nature structural & molecular biology*, 17(1), 133-138.

- Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5-6), 975-986.
- Kohli, R. M., Abrams, S. R., Gajula, K. S., Maul, R. W., Gearhart, P. J., & Stivers, J. T. (2009). A portable hot spot recognition loop transfers sequence preferences from APOBEC family members to activation-induced cytidine deaminase. *Journal of Biological Chemistry*, 284(34), 22898-22904.
- Koide, A., Wojcik, J., Gilbreth, R. N., Hoey, R. J., & Koide, S. (2012). Teaching an old scaffold new tricks: Monobodies constructed using alternative surfaces of the FN3 scaffold. *Journal of Molecular Biology*, 415(2), 393-405.
- Kontermann, R. E. (2011). Strategies for extended serum half-life of protein therapeutics. *Current Opinion in Biotechnology*, 22(6), 868-876.
- Kortemme, T., & Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22), 14116-14121.
- Kortemme, T., Joachimiak, L. a, Bullock, A. N., Schuler, A. D., Stoddard, B. L., & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nature structural & molecular biology*, 11(4), 371-379.
- Kortemme, T., Kim, D. E., & Baker, D. (2004). Computational alanine scanning of protein-protein interfaces. *Science's STKE : signal transduction knowledge environment*, 2004(219), pl2.
- Kortemme, T., Morozov, A. V., & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology*, 326(4), 1239-1259.
- Kortemme, T., Ramírez-Alvarado, M., & Serrano, L. (1998). Design of a 20-amino acid, three-stranded beta-sheet protein. *Science (New York, N.Y.)*, 281(5374), 253-256.
- Krehenbrink, M., Chami, M., Guilvout, I., Alzari, P. M., Pécorari, F., & Pugsley, A. P. (2008). Artificial binding proteins (Affitins) as probes for conformational changes in secretin PulD. *Journal of molecular biology*, 383(5), 1058-68.
- Kubala, M. H., Kovtun, O., Alexandrov, K., & Collins, B. M. (2010). Structural and thermodynamic analysis of the GFP:GFP-nanobody complex. *Protein science : a publication of the Protein Society*, 19(12), 2389-401.
- Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science (New York, N.Y.)*, 302(5649), 1364-1368.
- Kurochkina, L. P., & Mesyanzhinov, V. V. (1999). Co-expression of gene 31 and 23 products of bacteriophage T4. *Biochemistry. Biokhimiia*, 64(4), 379-383.
- Kuroda, D., Shirai, H., Jacobson, M. P., & Nakamura, H. (2012). Computer-aided antibody design. *Protein Engineering, Design and Selection*, 25(10), 507-521.
- Lamazares, E., Clemente, I., Bueno, M., Velázquez-Campoy, A., & Sancho, J. (2015). Rational stabilization of complex proteins: a divide and combine approach. *Scientific reports*, 5, 9129.

-
- Lanci, C., & MacDermaid, C. (2012). Computational design of a protein crystal. *Proceedings of the ...*, 109(19), 7304-7309.
- Lapidoth, G. D., Baran, D., Pszolla, G. M., Norn, C., Alon, A., Tyka, M. D., & Fleishman, S. J. (2015). AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins*, 83(8), 1385-406.
- Laskowski, R. a., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L., & Thornton, J. M. (1997). PDBsum: A Web-based database of summaries and analyses of all PDB structures. *Trends in Biochemical Sciences*, 22(12), 488-490.
- Le Gall, T., Romero, P. R., Cortese, M. S., Uversky, V. N., & Dunker, a K. (2007). Intrinsic disorder in the Protein Data Bank. *Journal of biomolecular structure & dynamics*, 24(4), 325-342.
- Leaver-Fay, A., Kuhlman, B., & Snoeyink, J. (2005). An adaptive dynamic programming algorithm for the side chain placement problem. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 27, 16-27.
- Leaver-Fay, A., O'Meara, M. J., Tyka, M., Jacak, R., Song, Y., Kellogg, E. H., ... Kuhlman, B. (2013). Scientific benchmarks for guiding macromolecular energy function improvement. *Methods in Enzymology*, 523, 109-143.
- Leaver-Fay, A., Snoeyink, J., & Kuhlman, B. (2008). On-the-fly rotamer pair energy evaluation in protein design. *Lecture Notes in Computer Science*, 4983 LNBI, 343-354.
- Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., ... Bradley, P. (2011). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487(C), 545-574.
- Lee, J. Il, Cho, S. S., Kil, E.-J., & Kwon, S.-T. (2010). Characterization and PCR application of a thermostable DNA polymerase from *Thermococcus pacificus*. *Enzyme and Microbial Technology*, 47(4), 147-152.
- Lee, J.-H., Kang, E., Lee, J., Kim, J., Lee, K. H., Han, J., ... Lee, J. Il. (2014). Protein grafting of p53TAD onto a leucine zipper scaffold generates a potent HDM dual inhibitor. *Nat Commun*, 5(May), 3815.
- Lemercier, G., Bakalara, N., & Santarelli, X. (2003). On-column refolding of an insoluble histidine tag recombinant exopolyphosphatase from *Trypanosoma brucei* overexpressed in *Escherichia coli*. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 786(1-2), 305-309.
- Lensink, M. F., & Wodak, S. J. (2010). Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins: Structure, Function and Bioinformatics*, 78(15), 3085-3095.
- Levy, R., Weiss, R., Chen, G., Iverson, B. L., & Georgiou, G. (2001). Production of correctly folded Fab antibody fragment in the cytoplasm of *Escherichia coli* trxB gor mutants via the coexpression of molecular chaperones. *Protein expression and purification*, 23(2), 338-347.
- Li, B., Zhao, L., Wang, C., Guo, H., Wu, L., Zhang, X., ... Guo, Y. (2010). The protein-protein interface evolution acts in a similar way to antibody affinity maturation. *Journal of Biological Chemistry*, 285(6), 3865-3871.
-

- Li, J., Lundberg, E., Vernet, E., Larsson, B., Höidén-Guthenberg, I., & Gråslund, T. (2010). Selection of affibody molecules to the ligand-binding site of the insulin-like growth factor-1 receptor. *Biotechnology and applied biochemistry*, 55(2), 99-109.
- Li, M., Jurado, K. a, Lin, S., Engelman, A., & Craigie, R. (2014). Engineered hyperactive integrase for concerted HIV-1 DNA integration. *PLoS one*, 9(8), e105078.
- Li, S. C., & Ng, Y. K. (2010). Calibur: a tool for clustering large numbers of protein decoys. *BMC bioinformatics*, 11, 25.
- Li, S., Schmitz, K. R., Jeffrey, P. D., Wiltzius, J. J. W., Kussie, P., & Ferguson, K. M. (2005). Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer cell*, 7(4), 301-11.
- Li, T., Pantazes, R. J., & Maranas, C. D. (2014). OptMAVEN – A New Framework for the de novo Design of Antibody Variable Region Models Targeting Specific Antigen Epitopes. *PLoS ONE*, 9(8), e105954.
- Liang, W. C., Wu, X., Peale, F. V., Lee, C. V., Meng, Y. G., Gutierrez, J., ... Fuh, G. (2006). Cross-species vascular endothelial growth factor (VEGF)-blocking antibodies completely inhibit the growth of human tumor xenografts and measure the contribution of stromal VEGF. *Journal of Biological Chemistry*, 281(2), 951-961.
- Linsley, P. S., & Ledbetter, J. A. (1993). The role of the CD28 receptor during T cell responses to antigen. *Annual review of immunology*, 191-212.
- Lippow, S. M., Wittrup, K. D., & Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. *Nature biotechnology*, 25(10), 1171-1176.
- Liu, S., Liu, S., Zhu, X., Liang, H., Cao, A., Chang, Z., & Lai, L. (2007). Nonnatural protein-protein interaction-pair design by key residues grafting. *Proceedings of the National Academy of Sciences of the United States of America*, 104(13), 5330-5335.
- Liu, Y., & Kuhlman, B. (2006). RosettaDesign server for protein design. *Nucleic Acids Research*, 34(WEB. SERV. ISS.), 235-238.
- Liu, Y., Zhu, Y., Ye, S., & Zhang, R. (2012). Crystal structure of kindlin-2 PH domain reveals a conformational transition for its membrane anchoring and regulation of integrin activation. *Protein & cell*, 3(6), 434-40.
- London, N., & Ambroggio, X. (2013). An accurate binding interaction model in de novo computational protein design of interactions: If you build it, they will bind. *Journal of Structural Biology*, 1-11.
- London, N., Gullá, S., Keating, A. E., & Schueler-Furman, O. (2012). In silico and in vitro elucidation of BH3 binding specificity toward Bcl-2. *Biochemistry*, 51(29), 5841-5850.
- Luo, L., Luo, Q., Guo, L., Lv, M., Lin, Z., Geng, J., ... Feng, J. (2014). Structure-based affinity maturation of a chimeric anti-ricin antibody C4C13. *Journal of biomolecular structure & dynamics*, 32(3), 416-23.

- Lyskov, S., Chou, F. C., Conchúir, S. Ó., Der, B. S., Drew, K., Kuroda, D., ... Das, R. (2013). Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLoS ONE*, 8(5), 5-7.
- Magis, C., Gasparini, D., Lecoq, a, Le Du, M., Stura, E., Charbonnier, J., ... Others. (2006). Structure-based secondary structure-independent approach to design protein ligands: application to the design of Kv1. 2 potassium channel blockers. *Journal of the American Chemical Society*, 128(50), 16190-16205.
- Mahajan, S., de Brevern, A. G., Sanejouand, Y.-H., Srinivasan, N., & Offmann, B. (2015). Use of a structural alphabet to find compatible folds for amino acid sequences. *Protein science : a publication of the Protein Society*, 24(1), 145-53.
- Mallajosyula, V. V. A., Citron, M., Ferrara, F., Temperton, N. J., Liang, X., Flynn, J. a., & Varadarajan, R. (2015). Hemagglutinin Sequence Conservation Guided Stem Immunogen Design from Influenza A H3 Subtype. *Frontiers in immunology*, 6(June), 329.
- Martin, F., Steinkühler, C., Brunetti, M., Pessi, a, Cortese, R., De Francesco, R., & Sollazzo, M. (1999). A loop-mimetic inhibitor of the HCV-NS3 protease derived from a minibody. *Protein engineering*, 12(11), 1005-1011.
- McCoy, a J., Chandana Epa, V., & Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces. *Journal of molecular biology*, 268(2), 570-584.
- Mei, B., Pan, C., Jiang, H., Tjandra, H., Strauss, J., Chen, Y., ... Murphy, J. E. (2010). Rational design of a fully active, long-acting PEGylated factor VIII for hemophilia A treatment. *Blood*, 116(2), 270-279.
- Méndez, R., Leplae, R., Lensink, M. F., & Wodak, S. J. (2005). Assessment of CAPRI predictions in Rounds 3-5 shows progress in docking procedures. *Proteins: Structure, Function and Genetics*, 60(2), 150-169.
- Miklos, A. E., Kluwe, C., Der, B. S., Pai, S., Sircar, A., Hughes, R. a., ... Ellington, A. D. (2012). Structure-based design of supercharged, highly thermoresistant antibodies. *Chemistry and Biology*, 19(4), 449-455.
- Miranda, F. F., Brient-Litzler, E., Zidane, N., Pecorari, F., & Bedouelle, H. (2011). Reagentless fluorescent biosensors from artificial families of antigen binding proteins. *Biosensors & bioelectronics*, 26(10), 4184-90.
- Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8), 114-117.
- Moran, N. (2011). Boehringer splashes out on bispecific antibody platforms. *Nature biotechnology*, 29(1), 5-6.
- Mouratou, B., Béhar, G., Paillard-Laurance, L., Colinet, S., & Pecorari, F. (2012). Ribosome display for the selection of Sac7d scaffolds. *Methods*, 805, 315-31.
- Mouratou, B., Schaeffer, F., Guilvout, I., Tello-Manigne, D., Pugsley, A. P., Alzari, P. M., & Pecorari, F. (2007). Remodeling a DNA-binding protein as a specific in vivo inhibitor of bacterial secretin PulD. *Proceedings of the National Academy of Sciences of the United States of America*, 104(46), 17983-8.

- Müller, H. N., & Skerra, a. (1994). Grafting of a high-affinity Zn(II)-binding site on the beta-barrel of retinol-binding protein results in enhanced folding stability and enables simplified purification. *Biochemistry*, 33(47), 14126-14135.
- Murphy, P. M., Bolduc, J. M., Gallaher, J. L., Stoddard, B. L., & Baker, D. (2009). Alteration of enzyme specificity by computational loop remodeling and design. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), 9215-9220.
- Murzin, A. G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *The EMBO journal*, 12(3), 861-7.
- Nelson, A. L., & Reichert, J. M. (2009). Development trends for therapeutic antibody fragments. *Nature biotechnology*, 27(4), 331-7.
- Nicaise, M., Valerio-Lepiniec, M., Minard, P., & Desmadril, M. (2004). Affinity transfer by CDR grafting on a nonimmunoglobulin scaffold. *Protein science : a publication of the Protein Society*, 13(7), 1882-1891.
- Nivón, L. G., Bjelic, S., King, C., & Baker, D. (2013). Automating human intuition for protein design. *Proteins*.
- Nivón, L. G., Moretti, R., & Baker, D. (2013). A Pareto-Optimal Refinement Method for Protein Design Scaffolds. *PLoS ONE*, 8(4), 1-5.
- Nygren, P. (2008). Alternative binding proteins: Affibody binding proteins developed from a small three-helix bundle scaffold. *Febs Journal*, 275, 2668-2676.
- O'Meara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., DiMaio, F., ... Kuhlman, B. (2015). Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *Journal of Chemical Theory and Computation*, 11(2), 609-622.
- Ochoa-Leyva, A., Soberón, X., Sánchez, F., Argüello, M., Montero-Morán, G., & Saab-Rincón, G. (2009). Protein Design through Systematic Catalytic Loop Exchange in the (β/α)₈ Fold. *Journal of Molecular Biology*, 387(4), 949-964.
- Oppmann, B., Lesley, R., Blom, B., Timans, J. C., Xu, Y., Hunte, B., ... Kastelein, R. a. (2000). Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. *Immunity*, 13(5), 715-725.
- Oshiro, S., & Honda, S. (2014). Imparting albumin-binding affinity to a human protein by mimicking the contact surface of a bacterial binding protein. *ACS Chemical Biology*, 9(4), 1052-1060.
- Pacheco, S., Béhar, G., Maillason, M., Mouratou, B., & Pecorari, F. (2014). Affinity transfer to the archaeal extremophilic Sac7d protein by insertion of a CDR. *Protein engineering, design & selection : PEDS*, 27(10), 431-8.
- Packer, M. S., & Liu, D. R. (2015). Methods for the directed evolution of proteins. *Nature Reviews Genetics*, 16(7), 379-394.
- Pantazes, R. J., Grisewood, M. J., Li, T., Gifford, N. P., & Maranas, C. D. (2015). The Iterative Protein Redesign and Optimization (IPRO) Suite of Programs. *Journal of Computational Chemistry*, 36, 251-263.

- Pantazes, R. J., & Maranas, C. D. (2010). OptCDR: A general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Engineering, Design and Selection*, 23(11), 849-858.
- Pantazes, R. J., & Maranas, C. D. (2013). MAPs: a database of modular antibody parts for predicting tertiary structures and designing affinity matured antibodies. *BMC bioinformatics*, 14(1), 168.
- Pegram, M., & Ngo, D. (2006). Application and potential limitations of animal models utilized in the development of trastuzumab (Herceptin[®]): A case study. *Advanced Drug Delivery Reviews*, 58(5-6), 723-734.
- Peng, K., Obradovic, Z., & Vucetic, S. (2004). Exploring bias in the Protein Data Bank using contrast classifiers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 435-446.
- Perticaroli, S., Nickels, J. D., Ehlers, G., O'Neill, H., Zhang, Q., & Sokolov, A. P. (2013). Secondary structure and rigidity in model proteins. *Soft Matter*, 9(40), 9548.
- Perticaroli, S., Nickels, J. D., Ehlers, G., & Sokolov, A. P. (2014). Rigidity, secondary structure, and the universality of the boson peak in proteins. *Biophysical Journal*, 106(12), 2667-2674.
- Pierce, N. A., & Winfree, E. (2002). Protein design is NP-hard. *Protein engineering*, 15(10), 779-782.
- Pletneva, N. V., Pletnev, V. Z., Lukyanov, K. a, Gurskaya, N. G., Goryacheva, E. a, Martynov, V. I., ... Pletnev, S. (2010). Structural evidence for a dehydrated intermediate in green fluorescent protein chromophore biosynthesis. *The Journal of biological chemistry*, 285(21), 15978-84.
- Ppyun, H., Kim, I., Cho, S. S., Seo, K. J., Yoon, K., & Kwon, S.-T. (2012). Improved PCR performance using mutant Tpa-S DNA polymerases from the hyperthermophilic archaeon *Thermococcus pacificus*. *Journal of biotechnology*, 164(2), 363-70.
- Prel, N. (2015). Pipeline : antibody-mimetics in Inflammation, Oncology and much more... Consulté à l'adresse <http://www.affilogic.com/pipeline> (consulté le 16 août 2015)
- Procko, E., Berguig, G. Y., Shen, B. W., Song, Y., Frayo, S., Convertine, A. J., ... Baker, D. (2014). A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. *Cell*, 157(7), 1644-1656.
- Procko, E., Hedman, R., Hamilton, K., Seetharaman, J., Fleishman, S. J., Su, M., ... Baker, D. (2013). Computational design of a protein-based enzyme inhibitor. *Journal of Molecular Biology*, 425(18), 3563-3575.
- Qiao, C., Lv, M., Li, X., Geng, J., Li, Y., Zhang, J., ... Shen, B. (2012). Affinity maturation of antiHER2 monoclonal antibody MIL5 using an epitope-specific synthetic phage library by computational design. *Journal of Biomolecular Structure and Dynamics*, (August 2015), 1-11.
- Ramamurthy, V., Krystek, S. R., Bush, A., Wei, A., Emanuel, S. L., Das Gupta, R., ... Sheriff, S. (2012). Structures of adnectin/protein complexes reveal an expanded binding footprint. *Structure (London, England : 1993)*, 20(2), 259-69.
- Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., ... Baker, D. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function and Bioinformatics*, 77(SUPPL. 9), 89-99.

- Reetz, M. T., Carballeira, J. D., & Vogel, A. (2006). Iterative saturation mutagenesis on the basis of b factors as a strategy for increasing protein thermostability. *Angewandte Chemie - International Edition*, 45(46), 7745-7751.
- Regan, L., & Clarke, N. D. (1990). A tetrahedral Zinc(II)-binding site introduced into a designed protein. *Biochemistry*, 29(49), 10878-10883.
- Riechmann, L., Clark, M., Waldmann, H., & Winter, G. (1988). Reshaping human antibodies for therapy. *Nature*.
- Robinson, H., Gao, Y. G., McCrary, B. S., Edmondson, S. P., Shriver, J. W., & Wang, A. H. (1998). The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature*, 392(6672), 202-5.
- RosettaCommons.Org. (2015). About RosettaCommons. Consulté à l'adresse <https://www.rosettacommons.org/about> (consulté le 17 août 2015)
- Rossmann, M. G. (1989). The canyon hypothesis. Hiding the host cell receptor attachment site on a viral surface from immune surveillance. *Journal of Biological Chemistry*, 264(25), 14587-14590.
- Rothbauer, U., Zolghadr, K., Muyldermans, S., Schepers, A., Cardoso, M. C., & Leonhardt, H. (2008). A versatile nanotrap for biochemical and functional studies with fluorescent fusion proteins. *Molecular & cellular proteomics : MCP*, 7(2), 282-9.
- Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., ... Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192), 190-195.
- Ruigrok, V. J. B., Levisson, M., Eppink, M. H. M., Smidt, H., & van der Oost, J. (2011). Alternative affinity tools: more attractive than antibodies? *The Biochemical journal*, 436(1), 1-13.
- Rutledge, S. E., Volkman, H. M., & Schepartz, A. (2003). Molecular Recognition of Protein Surfaces: High Affinity Ligands for the CBP KIX Domain. *Journal of the American Chemical Society*, 125(47), 14336-14347.
- Salgado, E. N., Ambroggio, X. I., Brodin, J. D., Lewis, R. a, Kuhlman, B., & Tezcan, F. A. (2010). Metal templated design of protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 107(5), 1827-1832.
- Sammond, D. W., Bosch, D. E., Butterfoss, G. L., Purbeck, C., MacHius, M., Siderovski, D. P., & Kuhlman, B. (2011). Computational design of the sequence and structure of a protein-binding peptide. *Journal of the American Chemical Society*, 133(12), 4190-4192.
- Sammond, D. W., Eletr, Z. M., Purbeck, C., Kimple, R. J., Siderovski, D. P., & Kuhlman, B. (2007). Structure-based Protocol for Identifying Mutations that Enhance Protein-Protein Binding Affinities. *Journal of Molecular Biology*, 371(5), 1392-1404.
- Sammond, D. W., Eletr, Z. M., Purbeck, C., & Kuhlman, B. (2010). Computational design of second-site suppressor mutations at protein-protein interfaces. *Proteins: Structure, Function and Bioinformatics*, 78(4), 1055-1065.

- Saraf, M. C., Moore, G. L., Goodey, N. M., Cao, V. Y., Benkovic, S. J., & Maranas, C. D. (2006). IPRO: an iterative computational protein library redesign and optimization procedure. *Biophysical journal*, 90(11), 4167-4180.
- Scalley-Kim, M., Minard, P., & Baker, D. (2003). Low free energy cost of very long loop insertions in proteins. *Protein science: a publication of the Protein Society*, 12(2), 197-206.
- Schaffitzel, C., Hanes, J., Jeremutis, L., & Plückthun, A. (1999). Ribosome display: an in vitro method for selection and evolution of antibodies from libraries. *Journal of immunological methods*, 231(1-2), 119-35.
- Schaffitzel, C., Zahnd, C., Amstutz, P., Luginbühl, B., & Plückthun, A. (2001). In Vitro Selection and Evolution of Protein-Ligand Interactions by Ribosome Display. In E. Golemis (Éd.), *Protein-Protein Interactions, A Molecular Cloning Manual* (p. 535-567). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Schein, C. H. (1989). Production of soluble recombinant proteins in bacteria. *Nature biotechnology*, 7, 1141-1149.
- Schmidt, M. M., & Wittrup, K. D. (2009). A modeling analysis of the effects of molecular size and binding affinity on tumor targeting. *Molecular cancer therapeutics*, 8(10), 2861-2871.
- Schönfeld, D., Matschiner, G., Chatwell, L., Trentmann, S., Gille, H., Hülsmeier, M., ... Skerra, A. (2009). An engineered lipocalin specific for CTLA-4 reveals a combining site with structural and conformational features similar to antibodies. *Proceedings of the National Academy of Sciences of the United States of America*, 106(20), 8198-203.
- Schueler-Furman, O., Wang, C., & Baker, D. (2005). Progress in protein-protein docking: Atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins: Structure, Function and Genetics*, 60(2), 187-194.
- Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Research*, 33(SUPPL. 2), 382-388.
- Sennhauser, G., Amstutz, P., Briand, C., Storchenegger, O., & Grütter, M. G. (2007). Drug export pathway of multidrug exporter AcrB revealed by DARPIn inhibitors. *PLoS Biology*, 5(1), 0106-0113.
- Sennhauser, G., & Grütter, M. G. (2008). Chaperone-Assisted Crystallography with DARPins. *Structure*, 16(10), 1443-1453.
- Sheffler, W., & Baker, D. (2009). RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. *Protein Science*, 18(1), 229-239.
- Sheffler, W., & Baker, D. (2010). RosettaHoles2: A volumetric packing measure for protein structure refinement and validation. *Protein Science*, 19(10), 1991-1995.
- Sia, S. K., & Kim, P. S. (2003). Protein grafting of an HIV-1-inhibiting epitope. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17), 9756-9761.
- Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L., ... Baker, D. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science (New York, N.Y.)*, 329(5989), 309-313.

- Simons, K. T., Bonneau, R., Ruczinski, I., & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Structure, Function and Genetics*, 37(SUPPL. 3), 171-176.
- Simons, K. T., Kooperberg, C., Huang, E., & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology*, 268(1), 209-225.
- Simonson, T., Gaillard, T., Mignon, D., Schmidt Am Busch, M., Lopes, A., Amara, N., ... Archontis, G. (2013). Computational protein design: The proteus software and selected applications. *Journal of Computational Chemistry*, 34(28), 2472-2484.
- Sircar, A., & Gray, J. J. (2010). SnugDock: Paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Computational Biology*, 6(1).
- Sivasubramanian, A., Chao, G., Pressler, H. M., Wittrup, K. D., & Gray, J. J. (2006). Structural model of the mAb 806-EGFR complex using computational docking followed by computational and experimental mutagenesis. *Structure*, 14(3), 401-414.
- Sivasubramanian, A., Sircar, A., Chaudhury, S., & Gray, J. J. (2009). Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins: Structure, Function and Bioinformatics*, 74(2), 497-514.
- Skerra, A. (2007). Alternative non-antibody scaffolds for molecular recognition. *Current opinion in biotechnology*, 18(4), 295-304.
- Skottrup, P. D. (2010). Small biomolecular scaffolds for improved biosensor performance. *Analytical Biochemistry*, 406(1), 1-7.
- Škrlec, K., Štrukelj, B., & Berlec, A. (2015). Non-immunoglobulin scaffolds: a focus on their targets. *Trends in biotechnology*, 33(7).
- 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), 1315-7.
- Steinmeyer, D. E., & McCormick, E. L. (2008). The art of antibody process development. *Drug discovery today*, 13(13-14), 613-8.
- Stranges, P. B., & Kuhlman, B. (2013). A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Science*, 22(1), 74-82.
- Stranges, P. B., Machius, M., Miley, M. J., Tripathy, A., & Kuhlman, B. (2011). Computational design of a symmetric homodimer using Beta-strand assembly. *Proceedings of the National Academy of Sciences*.
- Studier, F. W. (2005). Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification*, 41(1), 207-234.
- Su, S., Gao, Y. G., Robinson, H., Liaw, Y. C., Edmondson, S. P., Shriver, J. W., & Wang, a H. (2000). Crystal structures of the chromosomal proteins Sso7d/Sac7d bound to DNA containing T-G mismatched base-pairs. *Journal of molecular biology*, 303(3), 395-403.

- Subbotina, J., Yarov-Yarovoy, V., Lees-Miller, J., Durdagi, S., Guo, J., Duff, H. J., & Noskov, S. Y. (2010). Structural refinement of the hERG1 pore and voltage-sensing domains with ROSETTA-membrane and molecular dynamics simulations. *Proteins: Structure, Function and Bioinformatics*, 78(14), 2922-2934.
- Sudarshan, S., Kodathala, S. B., Mahadik, A. C., Mehta, I., & Beck, B. W. (2014). Protein-protein interface detection using the Energy Centrality Relationship (ECR) characteristic of proteins. *PLoS ONE*, 9(5).
- Sui, J., Hwang, W. C., Perez, S., Wei, G., Aird, D., Chen, L., ... Marasco, W. a. (2009). Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nature structural & molecular biology*, 16(3), 265-273.
- Sunada, H., Magun, B. E., Mendelsohn, J., & MacLeod, C. L. (1986). Monoclonal antibody against epidermal growth factor receptor is internalized without stimulating receptor phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, 83(11), 3825-9.
- Taylor, T. J., Bai, H., Tai, C. H., & Lee, B. (2014). Assessment of CASP10 contact-assisted predictions. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2), 84-97.
- Teze, D., Hendrickx, J., Dion, M., Tellier, C., Woods, V. L., Tran, V., & Sanejouand, Y. H. (2013). Conserved water molecules in family 1 glycosidases: A DXMS and molecular dynamics study. *Biochemistry*, 52(34), 5900-5910.
- Tlatli, R., Nozach, H., Collet, G., Beau, F., Vera, L., Stura, E., ... Cuniasse, P. (2013). Grafting of functional motifs onto protein scaffolds identified by PDB screening - An efficient route to design optimizable protein binders. *FEBS Journal*, 280(1), 139-159.
- Tokuriki, N., Tokuriki, N., Stricher, F., Stricher, F., Serrano, L., Serrano, L., ... Tawfik, D. S. (2008). How protein stability and new functions trade off. *PLoS Comput Biol*.
- Tornetta, M., Reddy, R., & Wheeler, J. C. (2012). Selection and maturation of antibodies by phage display through fusion to pIX. *Methods*.
- Trinquier, G., & Sanejouand, Y. H. (1998). Which effective property of amino acids is best preserved by the genetic code? *Protein engineering*, 11(3), 153-169.
- Tsai, C. J., Lin, S. L., Wolfson, H. J., & Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein science : a publication of the Protein Society*, 6(1), 53-64.
- Tsodikov, O. V., Thomas Record, M., & Sergeev, Y. V. (2002). Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *Journal of Computational Chemistry*.
- Umetsu, M., Nakanishi, T., Asano, R., Hattori, T., & Kumagai, I. (2010). Protein-protein interactions and selection: generation of molecule-binding proteins on the basis of tertiary structural information. *The FEBS journal*, 277(9), 2006-14.
- Van den Ent, F., & Löwe, J. (2006). RF cloning: a restriction-free method for inserting target genes into plasmids. *Journal of biochemical and biophysical methods*, 67(1), 67-74.

- Varga, I., Poczai, P., Cernák, I., & Hyvönen, J. (2014). Application of direct PCR in rapid rDNA ITS haplotype determination of the hyperparasitic fungus *Sphaeropsis visci* (Botryosphaeriaceae). *SpringerPlus*, 3(1880), 569.
- Vazquez, D. S., Agudelo, W. a., Yone, A., Vizioli, N., Arán, M., González Flecha, F. L., ... Santos, J. (2015). A helix-coil transition induced by the metal ion interaction with a grafted iron-binding site of the CyaY protein family. *Dalton Trans.*, 44(5), 2370-2379.
- Vihinen, M. (1987). Relationship of protein flexibility to thermostability. *Protein engineering*, 1(6), 477-480.
- Vincke, C., Loris, R., Saerens, D., Martinez-Rodriguez, S., Muyldermans, S., & Conrath, K. (2009). General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold. *Journal of Biological Chemistry*, 284(5), 3273-3284.
- Vita, C. (1997). Engineering novel proteins by transfer of active sites to natural scaffolds. *Current Opinion in Biotechnology*, 8(4), 429-434.
- Vita, C., Roumestand, C., Toma, F., & Ménez, A. (1995). Scorpion toxins as natural scaffolds for protein engineering. *Proceedings of the National Academy of Sciences of the United States of America*, 92(14), 6404-6408.
- Vita, C., Vizzavona, J., Drakopoulou, E., Zinn-Justin, S., Gilquin, B., & Menez, A. (1998). Novel miniproteins engineered by the transfer of active sites to small natural scaffolds. *Biopolymers - Peptide Science Section*.
- Voss, N. R., & Gerstein, M. (2010). 3V: Cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Research*.
- Vreven, T., Pierce, B. G., Hwang, H., & Weng, Z. (2013). Performance of ZDOCK in CAPRI rounds 20-26. *Proteins: Structure, Function and Bioinformatics*, 81(12), 2175-2182.
- Walser, R., Kleinschmidt, J. H., Skerra, A., & Zerbe, O. (2012). Beta-Barrel scaffolds for the grafting of extracellular loops from G-protein-coupled receptors. *Biological Chemistry*.
- Wang, C., Bradley, P., & Baker, D. (2007). Protein-Protein Docking with Backbone Flexibility. *Journal of Molecular Biology*, 373(2), 503-519.
- Wang, C., Vernon, R., Lange, O., Tyka, M., & Baker, D. (2010). Prediction of structures of zinc-binding proteins through explicit modeling of metal coordination geometry. *Protein Science*, 19(3), 494-506.
- Wang, Y. Sen, Youngster, S., Grace, M., Bausch, J., Bordens, R., & Wyss, D. F. (2002). Structural and biological characterization of pegylated recombinant interferon alpha-2b and its therapeutic implications. *Advanced Drug Delivery Reviews*, 54(4), 547-570.
- Wang, Y., Prosen, D. E., Mei, L., Sullivan, J. C., Finney, M., & Vander Horn, P. B. (2004). A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic acids research*, 32(3), 1197-207.
- Warwicker, J., Charonis, S., & Curtis, R. a. (2014). Lysine and arginine content of proteins: Computational analysis suggests a new tool for solubility design. *Molecular Pharmaceutics*, 11(1), 294-303.

- Weiner, L. M., Surana, R., & Wang, S. (2010). Monoclonal antibodies: versatile platforms for cancer immunotherapy. *Nature reviews. Immunology*, 10(5), 317-27.
- Wekerle, T., & Grinyó, J. M. (2012). Belatacept: from rational design to clinical application. *Transplant international : official journal of the European Society for Organ Transplantation*, 25(2), 139-50.
- Werther, W. a, Gonzalez, T. N., Connor, S. J. O., McCabe, S., Chan, B., Hotaling, T., ... Prestal, L. C. (1996). Humanization of an Anti-lymphocyte Function-Associated Antigen (LFA)-I Monoclonal Antibody and Reengineering. *Journal of immunology*, 24(27), 4986 - 4995.
- Whitehead, T. a, Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S. J., De Mattos, C., ... Baker, D. (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature Biotechnology*, 30(6), 543-548.
- Whitehead, T. a., Baker, D., & Fleishman, S. J. (2013). *Computational design of novel protein binders and experimental affinity maturation. Methods in Enzymology* (1^{re} éd., Vol. 523). Elsevier Inc.
- Wibley, J. E., Pegg, a E., & Moody, P. C. (2000). Crystal structure of the human O(6)-alkylguanine-DNA alkyltransferase. *Nucleic acids research*, 28(2), 393-401.
- Winter, G., Griffiths, A. D., Hawkins, R. E., & Hoogenboom, H. R. (1994). Making antibodies by phage display technology. *Annual review of immunology*, 12, 433-55.
- Wlodawer, A., Minor, W., Dauter, Z., & Jaskolski, M. (2013). Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS Journal*, 280(22), 5705-5736.
- Wörn, A., & Plückthun, A. (2001). Stability engineering of antibody single-chain Fv fragments. *Journal of molecular biology*, 305(5), 989-1010.
- Wu, S.-W., Ko, T.-P., Chou, C.-C., & Wang, A. H.-J. (2005). Design and characterization of a multimeric DNA binding protein using Sac7d and GCN4 as templates. *Proteins: Structure, Function, and Bioinformatics*, 60(4), 617-628.
- Xie, L., & Bourns, P. E. (2005). Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Computational Biology*, 1(3), 0222-0229.
- Xu, D., Lin, S. L., & Nussinov, R. (1997). Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *Journal of molecular biology*, 265(1), 68-84.
- Yang, J.-M., & Wang, A. H.-J. (2004). Engineering a Thermostable Protein with Two DNA-binding Domains Using the Hyperthermophile Protein Sac7d. *Journal of Biomolecular Structure and Dynamics*, 21(4), 513-526.
- Yarov-Yarovoy, V., Schonbrun, J., & Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. *Proteins*, 62(4), 1010-1025.
- Ye, Y., Shealy, S., Lee, H.-W., Torshin, I., Harrison, R., & Yang, J. J. (2003). A grafting approach to obtain site-specific metal-binding properties of EF-hand proteins. *Protein engineering*, 16(6), 429-434.

Yokoyama, K. I., Kunio, O., Ohtsuka, T., Nakamura, N., Seguro, K., & Ejima, D. (2002). In vitro refolding process of urea-denatured microbial transglutaminase without pro-peptide sequence. *Protein Expression and Purification*.

Zanghellini, A. (2014). De novo computational enzyme design. *Current Opinion in Biotechnology*, 29(1), 132-138.

Annexes

Annexes

Annexe 1. Scripts

Annexe 1.1. queue.sh

```
#!/bin/sh
# queue.sh
declare -a test=$(cat ./mutations-list-libX.txt)
declare -a pids=("1azp" "1c8c" "1lpj" "1qnt" "4f7h")
for i in "${pids[@]}"
do
    n=0
    echo "$test" | while read line
    do
        n=$((n+1))
        tmpname=$(echo $line | awk '{print $1}')
        echo "[chemin vers les exécutables]/fixbb.linuxgccrelease -database
[chemin vers la base de données] -s [chemin vers structures PDB
préparées]$i.relaxed.pdb -constant_seed -jran 1111111 -nstruct 10 -resfile [chemin vers
les fichiers resfiles]/$tmpname.resfile-$i -ex1 -ex2 -out:overwrite -out:scorefile
[chemin vers le dossier de scores]$tmpname-$i.1.sc -out:path:pdb [chemin vers le
dossier de sortie PDB]/ -out:suffix .$tmpname.1 > '[chemin vers le dossier de
journalisation]/$i-$tmpname.1.log'"
    done
done
```

Annexe 1.2. scores.sh

```
#!/bin/sh
# scores.sh
declare -a test=$(ls [chemin vers les scores]/)
echo -e "Sample Mean_score Standard_deviation"
echo "$test" | while read line
do
    liste=$(cat "[chemin vers les scores]/"$line | tail -n +3 | awk '{print $2}')
    somme=0
    iterations=$(echo "$liste" | wc -l)
    iterations=$(dc <<<"$iterations 0 + p")
    ecartype=0
    ecartypetmp=0
    while read i
    do
        somme=$(echo "scale=3;$somme+$i" | bc)
    done <<< "$liste"
    moyenne=$(echo "scale=3;$somme/$iterations" | bc)
    while read i
    do
        ecartypetmp=$(echo "$i - $moyenne" | bc)
        ecartypetmp=$(echo "scale=6; $ecartypetmp * $ecartypetmp" | bc)
        ecartype=$(echo "scale=6; $ecartype + $ecartypetmp" | bc)
    done <<< "$liste"
    ecartype=$(echo "scale=6;$ecartype/$iterations" | bc)
    ecartype=$(echo "scale=4;sqrt($ecartype)" | bc)
    echo -e `echo -e "$line\t$moyenne\t$ecartype" | awk -F '\t' '{gsub(/\t\./, "\t0.");
gsub(/\t-\./, "\t-0."); print $0}`
done
exit
```

Annexe 1.3. preparation.sh

Annexe 1.3.1. Script de préparation des fichiers resfile

```
#!/bin/sh
# preparation.sh
declare -a test=$(cat ./mutations-list-libX.txt)
declare -a resfiles=("resfile-1azp" "resfile-1c8c" "resfile-1lpj" "resfile-1qnt" "resfile-4f7h")
for i in "${resfiles[@]}"
do
    resfile=$(cat "./"$i)
    n=0
    echo "$test" | while read line
    do
        echo "$line"
        n=$((n+1))
        tmpname=$(echo $line | awk '{print $1}')
        tmp1=$(echo $line | awk '{print $2}')
        tmp2=$(echo $line | awk '{print $3}')
        tmp3=$(echo $line | awk '{print $4}')
        tmp4=$(echo $line | awk '{print $5}')
        tmp5=$(echo $line | awk '{print $6}')
        tmp6=$(echo $line | awk '{print $7}')
        tmp7=$(echo $line | awk '{print $8}')
        tmp8=$(echo $line | awk '{print $9}')
        tmp9=$(echo $line | awk '{print $10}')
        tmp10=$(echo $line | awk '{print $11}')
        tmp11=$(echo $line | awk '{print $12}')
        echo "$resfile" | awk '{gsub("%1%", "'$tmp1'"); gsub("%2%", "'$tmp2'");
gsub("%3%", "'$tmp3'"); gsub("%4%", "'$tmp4'"); gsub("%5%", "'$tmp5'"); gsub("%6%", "'$tmp6'");
gsub("%7%", "'$tmp7'"); gsub("%8%", "'$tmp8'"); gsub("%9%", "'$tmp9'");
gsub("%10%", "'$tmp10'"); gsub("%11%", "'$tmp11'"); print $0;}' > "./resfiles/"$tmpname"."$i
        done
    done
done
```

Annexe 1.3.2. Modèle pour 1AZP

```
NATAA # allow only the natural amino acid, this default command applies to all residues that
are not given non-default commands
start
#Lib X
21 A PIKAA %1% #Mutate the NF (chain A) with %1%
22 A PIKAA %2% #Mutate the NF (chain A) with %2%
24 A PIKAA %3% #Mutate the NF (chain A) with %3%
26 A PIKAA %4% #Mutate the NF (chain A) with %4%
29 A PIKAA %5% #Mutate the NF (chain A) with %5%
31 A PIKAA %6% #Mutate the NF (chain A) with %6%
33 A PIKAA %7% #Mutate the NF (chain A) with %7%
40 A PIKAA %8% #Mutate the NF (chain A) with %8%
42 A PIKAA %9% #Mutate the NF (chain A) with %9%
44 A PIKAA %10% #Mutate the NF (chain A) with %10%
46 A PIKAA %11% #Mutate the NF (chain A) with %11%
```

Annexe 1.3.3. Modèle pour 1C8C

```
NATAA # allow only the natural amino acid, this default command applies to all residues that
are not given non-default commands
start
#Lib X
21 A PIKAA %1% #Mutate the NF (chain A) with %1%
22 A PIKAA %2% #Mutate the NF (chain A) with %2%
24 A PIKAA %3% #Mutate the NF (chain A) with %3%
26 A PIKAA %4% #Mutate the NF (chain A) with %4%
29 A PIKAA %5% #Mutate the NF (chain A) with %5%
```

```

31 A PIKAA %6% #Mutate the NF (chain A) with %6%
33 A PIKAA %7% #Mutate the NF (chain A) with %7%
41 A PIKAA %8% #Mutate the NF (chain A) with %8%
43 A PIKAA %9% #Mutate the NF (chain A) with %9%
45 A PIKAA %10% #Mutate the NF (chain A) with %10%
47 A PIKAA %11% #Mutate the NF (chain A) with %11%

```

Annexe 1.3.4. Modèle pour 1LPJ

```

NATAA # allow only the natural amino acid, this default command applies to all residues that
are not given non-default commands
start
#Lib X
106 A PIKAA %1% #Mutate the NF (chain A) with %1%
107 A PIKAA %2% #Mutate the NF (chain A) with %2%
109 A PIKAA %3% #Mutate the NF (chain A) with %3%
111 A PIKAA %4% #Mutate the NF (chain A) with %4%
114 A PIKAA %5% #Mutate the NF (chain A) with %5%
116 A PIKAA %6% #Mutate the NF (chain A) with %6%
118 A PIKAA %7% #Mutate the NF (chain A) with %7%
125 A PIKAA %8% #Mutate the NF (chain A) with %8%
127 A PIKAA %9% #Mutate the NF (chain A) with %9%
129 A PIKAA %10% #Mutate the NF (chain A) with %10%
131 A PIKAA %11% #Mutate the NF (chain A) with %11%

```

Annexe 1.3.5. Modèle pour 1QNT

```

NATAA # allow only the natural amino acid, this default command applies to all residues that
are not given non-default commands
start
#Lib X
29 A PIKAA %1% #Mutate the NF (chain A) with %1%
30 A PIKAA %2% #Mutate the NF (chain A) with %2%
32 A PIKAA %3% #Mutate the NF (chain A) with %3%
34 A PIKAA %4% #Mutate the NF (chain A) with %4%
18 A PIKAA %5% #Mutate the NF (chain A) with %5%
20 A PIKAA %6% #Mutate the NF (chain A) with %6%
22 A PIKAA %7% #Mutate the NF (chain A) with %7%
7 A PIKAA %8% #Mutate the NF (chain A) with %8%
9 A PIKAA %9% #Mutate the NF (chain A) with %9%
11 A PIKAA %10% #Mutate the NF (chain A) with %10%
13 A PIKAA %11% #Mutate the NF (chain A) with %11%

```

Annexe 1.3.6. Modèle pour 4F7H

```

NATAA # allow only the natural amino acid, this default command applies to all residues that
are not given non-default commands
start
#Lib X
426 A PIKAA %1% #Mutate the NF (chain A) with %1%
427 A PIKAA %2% #Mutate the NF (chain A) with %2%
429 A PIKAA %3% #Mutate the NF (chain A) with %3%
431 A PIKAA %4% #Mutate the NF (chain A) with %4%
440 A PIKAA %5% #Mutate the NF (chain A) with %5%
442 A PIKAA %6% #Mutate the NF (chain A) with %6%
444 A PIKAA %7% #Mutate the NF (chain A) with %7%
451 A PIKAA %8% #Mutate the NF (chain A) with %8%
453 A PIKAA %9% #Mutate the NF (chain A) with %9%
455 A PIKAA %10% #Mutate the NF (chain A) with %10%
457 A PIKAA %11% #Mutate the NF (chain A) with %11%

```

Annexe 1.4. interactions-clusters.py

```

#!/usr/bin/python
# -*- coding: UTF-8 -*-

```

```

"""
Script: interactions-clusters.py
Auteur: Simon HUET, 2013, contact@simon-huet.fr
Liste des paramètres:
- i : fichier input = liste de fichiers pdb au format texte (1 chemin par ligne)
- o : output = dossier dans lequel les fichiers pdb créés seront stockés (mode 1) ou chemin
complet du fichier pdb (mode 2)
- c : cutoff pour la détermination des résidus à l'interface
- a : sélection dont on ne veut que les résidus à l'interface
- b : sélection à l'interface qui ne sera pas modifiée
- r : sélection qui doit être ignorée dans "a" dans le cas où on cherche le centre de
l'interaction
- m : mode du script
    -> 1 = considère le centre de (a - r) et génère 1 pdb par pose
    -> 2 = considère le centre de (a - r) et génère 1 pdb unique avec 1 état par pose
    -> 3 = considère la surface d'interaction de (a) et génère 1 pdb par pose
    -> 4 = considère la surface d'interaction de (a) et génère 1 pdb unique avec 1 état
par pose
    -> 5 = considère (a) entier et génère 1 pdb unique avec 1 état par pose
    -> 6 = considère le centre de la surface d'interaction de (a - r) et calcule la
distance maximum avec le centre

"""
# Paramètres par défaut
i = './test_input'
o = './output.pdb'
c = 4.0
a = 'chain A'
b = 'chain B'
r = 'resi 51-66'
m = 2

# Paramètres
argv = sys.argv[1:]
arg_count = 0
for arg in argv:
    arg = arg.split('=')
    if (arg[0] in('-h', '--help')) :
        print 'pymol -r interactions-clusters.py -- -i=<inputfile> -
o=<outputfile|outputdirectory> -c=<cutoff> -a=<chainA> -b=<chainB> -r=<removetoChainA> -
m=<mode>'
        sys.exit(0)
    else:
        arg_count += 1
        if arg_count == 1:
            print("Reading PDB structures from: %s\nSaving results in:
%s\nProcessing %s (not %s) + %s\nCutoff: %s\nMode: %s"%(i,o,a,r,b,str(c),str(m)))
            if arg[0] == '-i':
                i = arg[1].replace("'", "")
                print(i)
            elif arg[0] == '-o':
                o = arg[1].replace("'", "")
            elif arg[0] == '-c':
                c = float(arg[1].replace("'", ""))
            elif arg[0] == '-a':
                a = arg[1].replace("'", "")
            elif arg[0] == '-b':
                b = arg[1].replace("'", "")
            elif arg[0] == '-r':
                r = arg[1].replace("'", "")
            elif arg[0] == '-m':
                m = float(arg[1].replace("'", ""))

```

```

### Fonction interfaceResidues
from pymol import stored
def interfaceResidues(cmpx, cA='c. A', cB='c. B', cutoff=4.0, selName='interface'):
    """
    interfaceResidues -- finds 'interface' residues between two chains in a complex, from
    Jason Vertrees, 2009.
    """
    # Save user's settings, before setting dot_solvent
    oldDS = cmd.get("dot_solvent")
    cmd.set("dot_solvent", 1)

    # set some string names for temporary objects/selections
    tempC, selName1 = "tempComplex", selName+"1"
    chA, chB = "chA", "chB"

    # operate on a new object & turn off the original
    cmd.create(tempC, cmpx)
    cmd.disable(cmpx)

    # remove cruft and inrrelevant chains
    cmd.remove(tempC + " and not (polymer and (%s or %s))" % (cA, cB))

    # get the area of the complete complex
    cmd.get_area(tempC, load_b=1)
    # copy the areas from the loaded b to the q, field.
    cmd.alter(tempC, 'q=b')

    # extract the two chains and calc. the new area. Note: the q fields are copied to the
    new objects chA and chB
    cmd.extract(chA, tempC + " and (" + cA + ")")
    cmd.extract(chB, tempC + " and (" + cB + ")")
    cmd.get_area(chA, load_b=1)
    cmd.get_area(chB, load_b=1)

    # update the chain-only objects w/the difference
    cmd.alter( "%s or %s" % (chA,chB), "b=b-q" )

    # The calculations are done. Now, all we need to do is to determine which residues
    are over the cutoff and save them.
    stored.r = []
    cmd.iterate('%s or %s' % (chA, chB), 'stored.r.append((model,resi,b))')
    cmd.enable(cmpx)
    cmd.select(selName1, None)
    for (model,resi,diff) in stored.r:
        if abs(diff)>=float(cutoff):
            # expand the selection here
            cmd.select(selName1, selName1+" or "+model+" and resi "+resi)
    ## this is how you transfer a selection to another object.
    cmd.select(selName, cmpx + " in " + selName1)
    ## clean up after ourselves
    cmd.delete(selName1)
    cmd.delete(chA)
    cmd.delete(chB)
    cmd.delete(tempC)

    # reset users settings
    cmd.set("dot_solvent", oldDS)
    return True

cmd.extend("interfaceResidues", interfaceResidues)
### Fin fonction

### Début du run

```

```

print("Let's go!")
import time as tm # temps pour mesure du temps d'exécution
time1 = tm.time()
l_count = 0 # loop counter
cmd.delete('all') # delete everything
mean_dist = 0 # compteur de distance

# Lecture du fichier input (liste de fichiers pdb)
print("Reading input list.")
input_file = open(i,'r')
LIST = input_file.readlines()
input_file.close()

# Boucle pour chaque ligne
for line in LIST:
    l_count = l_count + 1
    print('Reading PDB #%. '%(str(l_count)))
    pdb = line.split()
    pdb = pdb[0] # élimine le problème de saut de ligne final
    cmd.load(pdb, 'complex') # load the pdb as 'complex'
    if( m in (1, 2, 3, 4, 6) ):
        interfaceResidues('complex', cutoff=c, selName='interface') # looking for
residues at interface
        if( m in (1, 2, 6) ):
            tmp_sel = '('+'a+' and '+'r+') or '
        elif( m in (3, 4) ):
            tmp_sel = ''
        cmd.remove('complex and ('+tmp_sel+'not (interface or '+b+'))') # remove
residues that are not at the interface on chain A
        if( m in (1, 2, 6) ):
            cmd.zoom('complex and %s'%a) # centre sur la chaîne A
            cmd.fragment('ala') # ajoute un fragment ala au centre de la zone de
contact
            cmd.set_name('ala', 'pa%s'%str(l_count)) # change le nom
            if( m == 6 ):
                # Calcul des distances
                myspace = {'tmp_dist': []}
                cmd.iterate('('+'a+' and n. CA)', 'tmp_dist.append(resi)',
space=myspace) # liste les numéros de résidus avec CA dans la chaîne A
                max_dist = 0
                for tmp_resi in myspace['tmp_dist']: # boucle sur les
résidus concernés
                    tmp_dist = cmd.distance('tmpd', 'pa%s and n.
CA'%str(l_count), 'resi %s and n. CA'%tmp_resi)
                    if(tmp_dist > max_dist):
                        max_dist = tmp_dist #stocke la plus grande
distance
                cmd.delete('all') # delete everything
                print('Maximum CA distance from interface center:
%s'%str(max_dist))
                mean_dist = mean_dist + max_dist
            else:
                cmd.remove('complex and %s'%a) # supprime la chaîne A
                if( m in (1, 3) ):
                    cmd.save(o+os.path.basename(pdb)) # saves the output pdb
                    cmd.delete('all') # delete everything
                if( m in (2, 4, 5) ):
                    if( m == 2 ):
                        tmp_sel = ' or pa%s'%str(l_count)
                    else:
                        tmp_sel = ''
                    if( m == 5 ):
                        int_sel = ''

```



```

else:
    int_sel = 'interface or '
    cmd.create('states', 'complex'+tmp_sel, 1, l_count)
    cmd.delete(int_sel+'complex'+tmp_sel) # delete temp structures

#Post boucle
if( m in (2, 4, 5) ):
    cmd.save(o, 'all', 0) # saves the output pdb
elif( m == 6):
    print('Mean radius: %s (Angstroms)'%(mean_dist/l_count))
### Fini !
print '\n Successful run!'
time2 = tm.time()
duration = time2 - time1
print ' Duration: %.2f second(s)'%duration

```

Annexe 1.5. pymol-CB-rms.py

```

#!/usr/bin/env python
# -*- coding: UTF-8 -*-
"""
Script: pymol-CB-rms.py
Auteur: Simon HUET, 2013, contact@simon-huet.fr
Liste des paramètres:
- in_name: Fichier d'entrée contenant la liste des fichiers PDB à analyser
- out_name: Fichier de sortie listant les n° de résidu et le nombre de changements de
coordonnées de CB associé. Note: un fichier avec le suffixe '.pymol' est également généré,
fournissant les commandes à exécuter dans PyMOL pour visualiser la carte des points chauds à
la surface de la GFP
"""
### A LANCER DEPUIS PYMOL
import sys, getopt, time
time1 = time.time()

# Valeurs par défaut, à modifier si nécessaire
in_name='./pdb.top200_Isc_-10_-5'
out_name='./PyMOL-resi_modifications'
ref_name='./CB_NF_ref.pdb'

# ouverture du fichier input avec la liste des fichiers pdb
with open(in_name, 'r') as f:
    pdblast = [line.strip() for line in f]

# assay

# Chargement de la structure de référence (avec CB de la Nanofitine non complexée)
cmd.load(ref_name, 'ref')

coord = {}

resi_list =
[2,3,5,6,7,8,9,11,12,13,14,15,16,17,18,19,21,22,23,25,26,27,28,29,30,32,34,36,37,38,39,41,42
,43,44,45,46,47,48,49,50,52,53,54,55,56,57,58,59,60,61,62,63,64,68,69,70,71,72,73,74,75,76,7
7,78,79,80,81,82,83,84,85,86,87,88,89,90,92,93,94,95,96,97,98,99,100,101,102,103,105,106,107
,108,109,110,111,112,113,114,115,117,118,119,120,121,122,123,124,125,126,128,129,130,131,132
,133,135,136,137,139,140,141,142,143,144,145,146,147,148,149,150,151,152,153,154,155,156,157
,158,159,161,162,163,164,165,166,167,168,169,170,171,172,173,175,176,177,178,179,180,181,182
,183,184,185,186,187,188,190,192,193,194,195,196,197,198,199,200,201,202,203,204,205,206,207
,208,209,210,211,212,213,214,215,216,217,218,219,220,221,223,224,225,226,227] # résidus
analysés

compteur = 0
for pdbitem in (pdblast):

```

```

compteur += 1
print 'PDB #s'%(str(compteur))
cmd.load(pdbitem, 'tmp')
for resi in (resi_list):
    if compteur == 1:
        coord['resi%s'%(str(resi))] = [resi, 0]
        if cmd.get_model('ref and c. B and n. CB and resi %s'%(str(resi)),
1).get_coord_list() != cmd.get_model('tmp and c. B and n. CB and resi %s'%(str(resi)),
1).get_coord_list() :
            coord['resi%s'%(str(resi))][1] = coord['resi%s'%(str(resi))][1] + 1
        cmd.delete('tmp')

result = ''
pymol_cmd = 'alter (gfp and n. CA),b=0;\n'

for resi in resi_list:
    result += '%s\t%s\n'%(str(resi),str(coord['resi%s'%(str(resi))][1]))
    new_b = float(100*(float(coord['resi%s'%(str(resi))][1]))/len(pdblist))
    pymol_cmd += 'alter gfp and n. CA and resi %s, b=%s;\n'%(str(resi),str(new_b))
pymol_cmd += "delete ramp_obj;\n\
delete ca_obj;\n\
delete legend;\n\
spectrum b, blue_white_red, gfp and n. CA, -50, 50;\n\
create ca_obj, gfp and n. CA;\n\
ramp_new ramp_obj, ca_obj, [-50, 50], [-1, -1, 0];\n\
set surface_color, ramp_obj, gfp;\n\
disable ca_obj;\n\
disable ramp_obj;\n\
ramp_new legend, ca_obj, [0, 50], [white, red];\n\
zoom (all), complete=1;"

#print result
out_file = open(out_name,'w')
out_file.write(result)
out_file.close()
out_file = open(out_name+'.pymol','w')
out_file.write(pymol_cmd)
out_file.close()

# Fini !
print '\n Finished!'
time2 = time.time()
duration = time2 - time1
print ' Duration: %.2f second(s)%duration

```

Annexe 1.6. interactions-resn-to-clusters.py

```

#!/usr/bin/python
# -*- coding: UTF-8 -*-
"""
Script: interactions-resn-to-clusters.py
Auteur: Simon HUET, 2014, contact@simon-huet.fr
Liste des paramètres:
- a : sélection à conserver dans les résidus à l'interface
- c : distance maximale entre CA et pseudoatom du cluster
- i : fichier input = liste de fichiers pdb au format texte (1 chemin par ligne)
- o : output = fichier dans lequel sont stockés les scores
- e : épitope à cibler en particulier
"""

### Paramètres par défaut
a = 'c. A and resi 7+8+9+21+22+24+26+29+31+33+40+42+44+46' # Ex: 'c. A' = Chaîne de la NF,
'c. A and resi 7+8+9+21+22+24+26+29+31+33+40+42+44+46' = Bq Y

```

```

c = 10.0 # Distance maximale entre CA et pseudoatome du cluster
i = './test.in'
o = './test-aa-banque-y-cluster2.out'
e = 'c. B and resi 182 and n. CA' # Pseudo-atome constituant le centre de l'épitope. Ex: 'c.
B and resi 206 and n. CA' = cluster1 et 'c. B and resi 182 and n. CA' = cluster2

# Paramètres
argv = sys.argv[1:]
arg_count = 0
for arg in argv:
    arg = arg.split('=')
    if (arg[0] in('-h', '--help')) :
        print 'pymol -r interactions-resn-to-clusters.py -- -i=<inputfile> -
o=<outputfile> -c=<max_distance> -a=<chain_to_analyze>'
        sys.exit(0)
    else:
        arg_count += 1
        if arg[0] == '-i':
            i = arg[1].replace("'", "")
            print(i)
        elif arg[0] == '-o':
            o = arg[1].replace("'", "")
        elif arg[0] == '-c':
            c = float(arg[1].replace("'", ""))
        elif arg[0] == '-a':
            a = arg[1].replace("'", "")
        elif arg[0] == '-e':
            e = arg[1].replace("'", "")

### Fonction interfaceResidues
from pymol import stored
def interfaceResidues(cmpx, cA='c. A', cB='c. B', cutoff=4.0, selName='interface'):
    """
    interfaceResidues -- finds 'interface' residues between two chains in a complex, from
    Jason Vertrees, 2009.
    """
    # Save user's settings, before setting dot_solvent
    oldDS = cmd.get("dot_solvent")
    cmd.set("dot_solvent", 1)

    # set some string names for temporary objects/selections
    tempC, selName1 = "tempComplex", selName+"1"
    chA, chB = "chA", "chB"

    # operate on a new object & turn off the original
    cmd.create(tempC, cmpx)
    cmd.disable(cmpx)

    # remove cruff and inrrelevant chains
    cmd.remove(tempC + " and not (polymer and (%s or %s))" % (cA, cB))

    # get the area of the complete complex
    cmd.get_area(tempC, load_b=1)
    # copy the areas from the loaded b to the q, field.
    cmd.alter(tempC, 'q=b')

    # extract the two chains and calc. the new area. # note: the q fields are copied to
    the new objects chA and chB
    cmd.extract(chA, tempC + " and (" + cA + ")")
    cmd.extract(chB, tempC + " and (" + cB + ")")
    cmd.get_area(chA, load_b=1)
    cmd.get_area(chB, load_b=1)

```

```
# update the chain-only objects w/the difference
cmd.alter( "%s or %s" % (chA,chB), "b=b-q" )

# The calculations are done. Now, all we need to do is to determine which residues
are over the cutoff and save them.
stored.r = []

cmd.iterate('%s or %s' % (chA, chB), 'stored.r.append((model,resi,b))')

cmd.enable(cmpx)
cmd.select(selName1, None)
for (model,resi,diff) in stored.r:

    if abs(diff)>=float(cutoff):

        # expand the selection here
        cmd.select(selName1, selName1+" or "+model+" and resi "+resi)

## this is how you transfer a selection to another object.
cmd.select(selName, cmpx + " in " + selName1)
## clean up after ourselves
cmd.delete(selName1)
cmd.delete(chA)
cmd.delete(chB)
cmd.delete(tempC)

# reset users settings
cmd.set("dot_solvent", oldDS)
return True

cmd.extend("interfaceResidues", interfaceResidues)
### Fin fonction

### Début du run
print("Let's go!")
import time as tm # temps pour mesure du temps d'exécution
time1 = tm.time()
l_count = 0 # loop counter

# Lecture du fichier input (liste de fichiers pdb)
print("Reading input list.")
input_file = open(i,'r')
LIST = input_file.readlines()
input_file.close()

out_file = open(o,'w')
# Boucle pour chaque ligne
for line in LIST:
    l_count = l_count + 1
    print('Reading PDB #%.'%(str(l_count)))
    pdb = line.split()
    pdb = pdb[0] # élimine le problème de saut de ligne final
    cmd.load(pdb, 'complex') # load the pdb as 'complex'

    interfaceResidues('complex', cA=a, selName='interface') # looking for residues at
interface
# Calcul des distances
myspace = {'tmp_dist': []}
cmd.iterate('(interface and ('+a+') and n. CA)', 'tmp_dist.append(resi)',
space=myspace) # liste les numéros de résidus avec CA dans la sélection a
within_dist = 0
for tmp_resi in myspace['tmp_dist']: # boucle sur les résidus concernés
```

```

        tmp_dist = cmd.distance('tmpd', e, 'complex and '+a+' and resi %s and n.
CA'%tmp_resi)
        if(tmp_dist < c):
            within_dist = within_dist + 1 # compte le nombre de CA suffisamment
proche du centre du cluster
            cmd.delete('complex or tmpd') # delete complex
            print('Number of CA within %s Ang from cluster center: %s'%(str(c),str(within_dist)))
            out_file.write((' %s\t%s\n'%(pdb,str(within_dist)))

#Post boucle
out_file.close()

### Fini !
print '\n Successful run!'
time2 = tm.time()
duration = time2 - time1
print ' Duration: %.2f second(s)'%duration

```

Annexe 1.7. weblogo-loop.sh

```

#!/bin/bash

### Fichiers pour générer la liste des commandes à exécuter
# Script: weblogo-loop.sh
# Version: 20130717
# Auteur: Simon HUET, 2013, contact@simon-huet.fr
# Liste des paramètres:
# - i: fichier d'entrée, avec 1) chemin vers un fichier au format FASTA contenant les
séquences, 2) Le titre du logo de séquence, 3) le nom du fichier à générer sur chaque ligne
(séparés par des tabulations)
###

# Defaults
input='./test.input'

# Input parameters
echo "Starting : $0"
declare -a parameters
for p in "$@"
do
    param=$(echo $p | cut -d"=" -f1)
    value=$(echo $p | cut -d"=" -f2)
    parameters[$param]=$value
done

# Loading input parameters
echo -e "\tParameters:"
#
if [ -z ${parameters["i"]} ]
then
    echo -e "\t(input set to default)"
else
    input=${parameters["i"]}
fi
echo -e "\tinput: $input"
#

cat $input | while read line || [ -n "$line" ]; do
    echo "$line" | awk -F'\t' '{system("weblogo -F PNG -f " $1 " -o " $3 " -U probability -
i -222 -l 1 -u 66 -s large -n 70 -t \"\" $2 "\" --label \"\" -x \"Position and amino-acids 1
letter code\" -y \"Frequency\" --annotate
\"S,K,G,A,E,L,F,T,G,I,V,P,I,L,I,E,L,N,G,D,V,N,G,H,K,F,S,V,S,G,E,G,E,G,D,A,T,Y,G,K,L,T,L,K,F,
I,C,T,T,G,K,L,P,V,P,W,P,T,L,V,T,T,L,V,Q,C,F,S,R,Y,P,D,H,M,K,Q,H,D,F,F,K,S,A,M,P,E,G,Y,I,Q,E,

```

```
R,T,I,F,F,E,D,D,G,N,Y,K,S,R,A,E,V,K,F,E,G,D,T,L,V,N,R,I,E,L,T,G,T,D,F,K,E,D,G,N,I,L,G,N,K,M,
E,Y,N,Y,N,A,H,N,V,Y,I,M,T,D,K,A,K,N,G,I,K,V,N,F,K,I,R,H,N,I,E,D,G,S,V,Q,L,A,D,H,Y,Q,Q,N,T,P,
I,G,D,G,P,V,L,L,P,D,N,H,Y,L,S,T,Q,S,A,L,S,K,D,P,N,E,K,R,D,H,M,I,Y,F,G,F,V,T,A,A,M1,V2,K3,V4,
K5,F6,K7,Y8,K9,G10,E11,E12,K13,E14,V15,D16,T17,S18,K19,I20,K21,K22,V23,W24,R25,V26,G27,K28,M
29,V30,S31,F32,T33,Y34,D35,D36,N37,G38,K39,T40,G41,R42,G43,A44,V45,S46,E47,K48,D49,A50,P51,K
52,E53,L54,L55,D56,M57,L58,A59,R60,A61,E62,R63,E64,K65,K66\" -P \"\" -c chemistry --
errorbars NO"}}'
# Sequence annotation and res. range adjusted to GFP:Nanofitin
done
```

Annexe 1.8. `pdbsurf-auto.sh` (v1.3)

```
#!/bin/sh
#=====
# Script: pdbsurf-auto.sh
# Auteur: Simon Huet, 2014, contact@simon-huet.fr
#-----
#ChnG v1.0 : Version initiale basée sur la v1.15 de Pdbsurf.sh (YH Sanejouand, Avril 2013)
#ChnG v1.1 : Recherche de Surface Racer dans ~/Test/scripts/surfaceracer/ en priorité
#ChnG v1.2 : Compare un PDB en complexe AB et les PDB de A et B
#ChnG v1.3 : Ajout de 3v pour calcul de volume de cavité à l'interface
version='v1.3, May 2014.'

#Name of the executable:
racerexe=surfrace5_0_linux_64bit

#Name of the file with the radii:
rfile=radii.txt

if [ $# -gt 0 ]
then
    nfil=`ls -1d $* | wc -l`
else
    nfil=0
fi
if [ $nfil -eq 0 ]
then
    echo 'Usage: pdbsurf "pdb-file-selection"'
    echo 'Action: Accessible surface area for a set of PDB files, using Surface Racer.'
    exit
fi
prognm='.Pdbsurf>'
echo $prognm $version

echo $prognm 'Accessible surface area for a set of' $nfil 'PDB file(s).'

```

```

# Not perfect: the database for locate could need an update; slow otherwise.
racesurf=`locate $racerexe | tail -1`
if [ ! -x "$racesurf" -o ! -n "$racesurf" ]
then
echo $progr 'Surface Racer executable:' $racerexe 'not
found.'
exit
fi
fi
fi
cp $racesurf .
fi
echo $progrm 'Executable to be used locally:' $racesurf

if [ -f $rfile ]
then
echo $progrm 'Local radii file to be used:' $rfile
else
racedir=`dirname $racesurf`
rfilnam=$racedir/'$rfile'
if [ -f $rfilnam ]
then
echo $progrm 'Radii file:' $rfilnam 'copied locally.'
cp $rfilnam .
else
echo $progr 'Radii file:' $rfile 'not found in directory' $racedir
exit
fi
fi

savdir='output'
tstwd=`echo $savdir | tr '[a-z]' '[A-Z]`
echo "savedir test: $tstwd"
case $tstwd in
NONE|NON|NO|N) savdir=NONE ;;
*)
if [ ! -d $savdir ] ; then mkdir $savdir ; fi
echo $progrm 'Output PDB file(s) to be saved in directory:' $savdir ;;
esac

#CoM Surface Racer has sometimes problems, the symptom being a very low SA.
if [ ! $savdir = NONE ]
then
echo ' Minimum accessible surface expected for saving results ? (<0: Any)'
#read minsrf
minsrf=0
case $minsrf in *[-9,!\.]*) minsrf=-1 ;; esac
fi

# Input for Surface Racer:
# -----
# Radii set (1 or 2)
# PDB
# Probe (water) radius (1.4: valeur du papier)
# SA (1) SA+MSA (2) or SA+MSA+Curvature (3)
cat > surface_racer.inp_pdbsurf << EOFsr
1
structure.pdb
1.4
1
EOFsr

```

```
if test -f pbsurf.sum
then
  echo $progwn 'pbsurf.sum file already exists. A copy is saved.'
  mv pbsurf.sum pbsurf.sum_old
fi
echo $prognm $version > pbsurf.sum
echo $prognm $version > pbsurf-cav.sum
echo -e "PDB\Int_surf\Polar_int_surf\Nonpolar_int_surf\Int_cavity\Int_cav_volume" >
pbsurf-cav.sum

# =====
# Input files are scanned:
# =====
nlwsa=0
nfnf=0
nfb=0
for j in `ls -d $*` # all complex PDB
do
  # counters
  intcav=0
  intsurf=0
  intpolsurf=0
  intnpolsurf=0
  iloop=0

  for i in "$j" "$j.cA" "$j.cB" # for AB, A and B
  do
    iloop=$((iloop+1))
    if [ -f $i ] ; then
      echo $prognm 'File considered:' $i

      # Gzip eventual:
      if [ ${i##*.} = "gz" ]
      then
        zcat $i > structure.pdb
      else
        cp $i structure.pdb
      fi

      natom=`cut -c1-4 structure.pdb | egrep -c 'ATOM|HETA'`
      if [ $natom -le 0 ]
      then
        echo $progwn 'No line with ATOM or HETA keys, as expected in a PDB file.'
      else
        # =====
        # Surface Racer:
        # =====
        rm -f structure.txt structure_residue.txt result.txt pbsurf.out
        ./$racerexe < surface_racer.inp_pbsurf > pbsurf.out

        if [ -f structure.txt -a -f pbsurf.out ] ; then
          nok=`grep -A1 'Solvent accessible surface areas' pbsurf.out | grep -c`
        else
          nok=0
        fi

        # Calculation ended normally:
        if [ $nok -gt 0 ]
        then
          srout=`grep -A1 'Solvent accessible surface areas' pbsurf.out | tail -1`
        fi
      fi
    fi
  done
done
```



```

#       Nombre de cavites trouvees:
      nca=`cut -c1-6 structure.txt | grep -c CAVITY`
      echo $progrnm $ncav 'cavities found.'

#       Nombre de CA:
      nca=`grep 'ATOM' structure.txt | cut -c 12-17 | grep -c ' CA '`
      echo $i $natom 'atoms,' $nca 'aa,' $ncav 'cavities,' $srout | sed 's/
/,/,/g' >> pdsurf.sum

# Counting interface values
if [ $iloop -eq 1 ]
then
      intcav=$ncav
      intsurf=$(echo 0 - $(echo "$srout" | sed -n "s/.Total area = \([0-9]*\.[0-9]*\).*\1/p") | bc)
      intpolsurf=$(echo 0 - $(echo "$srout" | sed -n "s/.Polar area= \([0-9]*\.[0-9]*\).*\1/p") | bc)
      intnpolsurf=$(echo 0 - $(echo "$srout" | sed -n "s/.Non-polar area= \([0-9]*\.[0-9]*\).*\1/p") | bc)
      else
      intcav=$((intcav - $ncav))
      intsurf=$(echo $intsurf + $(echo "$srout" | sed -n "s/.Total area = \([0-9]*\.[0-9]*\).*\1/p") | bc)
      intpolsurf=$(echo $intpolsurf + $(echo "$srout" | sed -n "s/.Polar area= \([0-9]*\.[0-9]*\).*\1/p") | bc)
      intnpolsurf=$(echo $intnpolsurf + $(echo "$srout" | sed -n "s/.Non-polar area= \([0-9]*\.[0-9]*\).*\1/p") | bc)
      fi

#       =====
#       Sauvegarde:
#       =====
      if [ ! $savdir = NONE ]
      then
      satot=`echo $srout | cut -f2 -d=' ' | cut -f1 -d',' | sed 's/ //g'`
      if [ `echo "$satot > $minsrf" | bc` -eq 1 ]
      then
      filnam=`basename $i`

#       On extrait l'info sur les cavites:
      if [ $ncav -gt 0 ]
      then
#       Numero de la premiere ligne avec la cle CAVITY:
      nfrst=`cut -c1-6 structure.txt | grep -n CAVITY | head -1 | cut -f1
-d': '`

      nfrst=`expr $nfrst - 1`
      nlines=`wc -l structure.txt | cut -f1 -d' '`
      nlast=`expr $nlines - $nfrst`

      head -$nfrst structure.txt > structure_nocav.txt
      tail -$nlast structure.txt > structure_cav.txt
      mv structure_nocav.txt structure.txt
      mv structure_cav.txt $savdir/$filnam'_surfcav'

      fi

#       On garde l'entete du fichier PDB:
      nfrst=`cut -c1-4 structure.pdb | egrep -n 'ATOM|HETA' | head -1 | cut -f1
-d': '`

      nfrst=`expr $nfrst - 1`

#       Output in true PDB format, with ASA instead of B-factors:
#       (Surface Racer gives radius, ASA, MSA, Curvature, after x,y,z)
      cut -c1-54 structure.txt > structure_xyz.txt
#       Alignement des virgules:

```

```

cut -c61-68 structure.txt | awk '{ if ( $1 < 10 ) print " " $1 ; else
print $1 }' > structure_surf.txt

head -nfrst structure.pdb > structure.txt
paste -d';' structure_xyz.txt structure_surf.txt | sed 's/;/ /' >>
structure.txt

echo END >> structure.txt
mv structure.txt $savdir/'$filnam'_surf'

# Surface minimum respectee:
else
nlowsa=`expr $nlowsa + 1`
fi

# Sauvegardes demandees:
fi

nfnfnd=`expr $nfnfnd + 1`
lastok=$i
else
echo $proger 'Surface Racer failed.'
tail -10 pdbsurf.out
echo '...'
echo $proger 'Surface Racer failed for file' $i > pdbsurf.out_lastpb
echo $proger 'Surface Racer output:' >> pdbsurf.out_lastpb
cat pdbsurf.out >> pdbsurf.out_lastpb
npb=`expr $npb + 1`
lastpb=$i
fi

# PDB file looks fine:
fi
# File exists:
fi
done
echo $progrnm 'Running 3V for cavity volume'
# Recuperation des atomes des cavites uniquement à l'interface
filnam=`basename $j`
grep -f $savdir/'$filnam'.cA_surfcav' -v $savdir/'$filnam'_surfcav' | grep -f
$savdir/'$filnam'.cB_surfcav' -v > diff.pdb
# Conversion et calcul de volume de cavite
vpath='/comptes/E072042G/Test/2014/20140520-test-3v/3v/'
if [ ! -f 'atmtypenumbers' ]
then
cp $vpath'xyzr/atmtypenumbers' ./
fi
$vpath/xyzr/pdb_to_xyzr ./diff.pdb > ./diff.xyzr
$vpath/bin/Volume.exe -i ./diff.xyzr -p 3 -g 0.5 > ./diff.txt
cavol=`cat ./diff.txt | cut -f 3`
intsurf=$(echo 'scale=2;' $intsurf/2 | bc)
intpolsurf=$(echo 'scale=2;' $intpolsurf/2 | bc)
intnpolsurf=$(echo 'scale=2;' $intnpolsurf/2 | bc)
echo -e "$j\t$intsurf\t$intpolsurf\t$intnpolsurf\t$intcav\t$cavol" >> pdbsurf-cav.sum

done
if [ ! $savdir = NONE ] ; then cp pdbsurf.sum $savdir/pdbsurf.sum_sav ; fi

case $nfnfnd in
0) echo $proger 'No successful surface calculation could be performed.' ; exit ;;
1) echo $progrnm 'Surface calculated for file:' $lastok ;;
*) echo $progrnm 'Surface calculated for' $nfnfnd 'files.' ;;
esac
if [ $nlowsa -gt 0 ] ; then echo $progrwn 'Surface lower than' $mnsrfr 'for' $nlowsa
'file(s).' ; fi

```

```

case $npb in
0) echo $progrm 'Everything went fine, it seems !' ;;
1) echo $progrm 'Surface could not be calculated for file:' $lastpb ;;
*) echo $progrm 'Surface could not be calculated for' $npb 'files.' ;;
esac

#Some cleaning:
rm -f structure.pdb structure_xyz.txt structure_surf.txt structure.txt structure_cav.txt
surface_racer.inp_pdbsurf
rm -f structure_residue.txt result.txt pdbsurf.out
rm -f diff.pdb diff.xyzz diff.txt atmtypenumbers diff_residue.txt diff_surfcav.pdb
diff.txtfcav.pdb
exit

```

Annexe 1.9. batch.sh

```

#!/bin/bash

### Génération de rounds de docking/relax/design
# Author: Simon Huet, contact@simon-huet.fr
# Date: 201504
# Current version: 1.3
# Changelog:
# 1.0: à partir d'une pose de départ, cycles de docking/design/relax avec extraction
#       topscores et postraitements tels que rms, sequence, i_sc etc.
# 1.1: + mock mode
#       + filtre séquences uniques et retrait des rms calculés lors du design (base sur CA
#       donc inutile dans ce cas)
#       + options spin/rand1/rand2
# 1.2: correction bugs lecture scores (cut de fichiers avec largeur fixes variables et grep
#       sur nom de fichier input non adaptés en batches)
# 1.3: input issu d'un tour précédent possible (fichier .sum)
###
###

# Env
export LC_NUMERIC="en_US.UTF-8" #Eviter problemes de separateurs de decimales

# Defaults
amsterdam=1 # sur amsterdam ou non
s='' # pose initiale
o='input.txt' # fichier texte dans lequel récupérer la liste, n'est pas écrasé mais
# continué, dans le répertoire "f"
f='~/Test/tmp' # répertoire de travail de rosetta
initial_offset=0 #offset de départ (nombre de jobs precedents)
simult_proc=1 #processeurs a utiliser en simultane
#decoys=1 # nombre de decoys par job
pert_deg=3 # angle de perturbation initiale, en degrés, dans docking
pert_dist=8 # distance de perturbation initiale, en angstrom, dans docking
spin=0 # bolean pour option -spin, dans docking
rand1=0 # bolean pour option -randomize1, dans docking
rand2=0 # bolean pour option -randomize2, dans docking
unitrans=0 # -uniform_trans value (ignored if 0 by default), dans docking
max=1000 # nombre max de decoys par job
rounds='relax:2:1,docking:2:1' # rounds à faire, syntaxe: 'relax:5:1,docking:10,design:10'
# etc. [type de round (relax|docking|design)]:[nombre de decoys à faire au total]:[nombre de
# decoys à garder en fin de round]
rmsmin=0 #valeur min
rmsmax=0.5 #valeur max
mock=0 # bolean, 1 si les calculs longs ne sont pas à lancer

```

```
# Input parameters
echo -e "[$(date +%Y-%m-%d:%H:%M:%S)] Starting:\n\t$0"

declare -A parameters

for p in "$@"
do
    #echo -e "\tParameters : $p"
    param=$(echo $p | cut -d"=" -f1)
    value=$(echo $p | cut -d"=" -f2)
    #echo -e "\t$param: $value"
    parameters[$param]=$value
    #echo -e "\t\t${parameters[$param]}"
done

#echo ${!parameters[*]}
#echo ${parameters[*]}

# Loading input parameters
echo -e "[$(date +%Y-%m-%d:%H:%M:%S)] Parameters:"
#
if [ -z ${parameters["amsterdam"]} ]
then
    echo -e "\t(amsterdam set to default)"
else
    amsterdam=${parameters["amsterdam"]}
fi
echo -e "\tamsterdam: $amsterdam"
if [ $amsterdam -eq 1 ]
then
    rosettapath='rosettamsterdam'
else
    rosettapath='rosetta'
fi
#
if [ -z ${parameters["s"]} ] # si la variable n'est pas déclarée
then
    echo -e "\tInitial pose required! (s parameter)"
    exit
else
    s=${parameters["s"]}
fi
echo -e "\ts: $s"
#
if [ -z ${parameters["o"]} ]
then
    echo -e "\t(o set to default)"
else
    o=${parameters["o"]}
fi
echo -e "\to: $o"
#
if [ -z ${parameters["f"]} ]
then
    echo -e "\t(f set to default)"
else
    f=${parameters["f"]}
fi
echo -e "\tf: $f"
#
if [ -z ${parameters["initial_offset"]} ]
then
    echo -e "\t(initial_offset set to default)"
```

```
else
    initial_offset=${parameters["initial_offset"]}
fi
echo -e "\tinitial_offset: $initial_offset"
#
if [ -z ${parameters["simult_proc"]} ]
then
    echo -e "\t(simult_proc set to default)"
else
    simult_proc=${parameters["simult_proc"]}
fi
echo -e "\tsimult_proc: $simult_proc"
#
#if [ -z ${parameters["decoys"]} ]
#then
#    echo -e "\t(decoys set to default)"
#else
#    decoys=${parameters["decoys"]}
#fi
#echo -e "\tdecoys: $decoys"
#
if [ -z ${parameters["pert_deg"]} ]
then
    echo -e "\t(pert_deg set to default)"
else
    pert_deg=${parameters["pert_deg"]}
fi
echo -e "\tpert_deg: $pert_deg"
#
if [ -z ${parameters["pert_dist"]} ]
then
    echo -e "\t(pert_dist set to default)"
else
    pert_dist=${parameters["pert_dist"]}
fi
echo -e "\tpert_dist: $pert_dist"
#
if [ -z ${parameters["spin"]} ]
then
    echo -e "\t(spin set to default)"
else
    spin=${parameters["spin"]}
fi
echo -e "\tspin: $spin"
#
if [ -z ${parameters["rand1"]} ]; then echo -e "\t(rand1 set to default)"; else
rand1=${parameters["rand1"]} ; fi
echo -e "\trand1: $rand1"
#
if [ -z ${parameters["rand2"]} ]; then echo -e "\t(rand2 set to default)"; else
rand2=${parameters["rand2"]} ; fi
echo -e "\trand2: $rand2"
#
if [ -z ${parameters["unitrans"]} ]; then echo -e "\t(unitrans set to default)"; else
unitrans=${parameters["unitrans"]} ; fi
echo -e "\tunitrans: $unitrans"
#
if [ -z ${parameters["max"]} ]
then
    echo -e "\t(max set to default)"
else
    max=${parameters["max"]}
fi
```

```

echo -e "\tmax: $max"
#
if [ -z ${parameters["rounds"]} ]
then
echo -e "\t(rounds set to default)"
else
rounds=${parameters["rounds"]}
fi
echo -e "\trounds: $rounds"
#
if [ -z ${parameters["rmsmin"]} ]; then echo -e "\t(rmsmin set to default)"; else
rmsmin=${parameters["rmsmin"]} ; fi
echo -e "\trmsmin: $rmsmin"
#
if [ -z ${parameters["rmsmax"]} ]; then echo -e "\t(rmsmax set to default)"; else
rmsmax=${parameters["rmsmax"]} ; fi
echo -e "\trmsmax: $rmsmax"
#
if [ -z ${parameters["mock"]} ]; then echo -e "\t(mock set to default)"; else
mock=${parameters["mock"]} ; fi
echo -e "\tmock: $mock"
#

# Creating directories
if [ ! -d $f ] ; then mkdir $f ; fi
if [ ! -d "$f/input" ] ; then mkdir "$f/input" ; fi
if [ ! -d "$f/scores" ] ; then mkdir "$f/scores" ; fi
if [ ! -d "$f/logs" ] ; then mkdir "$f/logs" ; fi
if [ ! -d "$f/output_pdb" ] ; then mkdir "$f/output_pdb" ; fi

# Copying input structure
cp $s "$f/input/"

# Rounds system
nround=0
IFS=', ' read -a array <<< "$rounds"
for round in "${array[@]}"
do # pour chaque round

    # Round en cours
    #paramètres
    type=$(echo $round | cut -d":" -f1) #type de round
    tot=$(echo $round | cut -d":" -f2) #total de decoys
    top=$(echo $round | cut -d":" -f3) #nombre de decoys à conserver
    nround=$((nround+1))

    echo -e "[$(date +%Y-%m-%d:%H:%M:%S)] Running round #${nround}:\tOverall $top top
results will be kept from the $tot decoys of $type round per input decoy (max $max decoys
given per job, $simult_proc simultaneous proc)"
    # Creating directories
    if [ ! -d $f ] ; then mkdir $f ; fi
    if [ ! -d "$f/input/$nround-$type" ] ; then mkdir "$f/input/$nround-$type" ; fi
    if [ ! -d "$f/scores/$nround-$type" ] ; then mkdir "$f/scores/$nround-$type" ; fi
    if [ ! -d "$f/logs/$nround-$type" ] ; then mkdir "$f/logs/$nround-$type" ; fi
    if [ ! -d "$f/output_pdb/$nround-$type" ] ; then mkdir "$f/output_pdb/$nround-$type"
; fi

    # Inputs
    inputs=$s
    # Verifie si input est un fichier .sum d'un tour précédent
    filename=$(basename "$inputs"); extension="${filename##*.}"; if [ "$extension" ==
'sum' ] ; then topsum="$inputs"; fi

```

```

if [ -z $topsum ] # Si $topsum non défini (donc pas de run précédent réussi dans la
loop)
then
    if [ $nround -eq 1 ] # Si tour = 1, alors pas besoin de $topsum défini
    then
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tNo previous round
used as input. Using starting pose: $s"
    else #Sinon il y a un problème
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tAn error occured! Previous
round top poses not defined!"
        exit
    fi
else
    if [ -f "$topsum" ] # Si fichier $topsum existe (donc run précédent réussi
dans la loop)
    then
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tTop poses list found as
input: $topsum"
        inputs=$(tail -n +2 $topsum)
    else
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tAn error occured! Previous
round top poses not found!"
        exit
    fi
fi
#
echo '' > "$f/input/$nround-$type/$o" # reset la commande
echo '' > "$f/input/$nround-$type/rmsd.$o" # reset la commande

# Cases
case "$type" in

# Relax
'relax') echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tStarting relax"
#
c=$((1+$initial_offset)) # compteur
# Liste inputs
echo -e "$inputs" | while read i
do
    input=$(echo -e "$i" | tr -s ' ' | cut -d' ' -f3) # récupère chaque
chemin complet de fichier
    echo "input vaut: $input"

# Boucle tant qu'il y a des decoys à faire
remainingjobs=$tot
while [ $remainingjobs -gt 0 ]
do
    seed=$((1111111+$c)) #random_seed pas random
# nombre de structures à faire
if [ $remainingjobs -lt $max ]; then nstruct=$remainingjobs;
else nstruct=$max; fi

#relax itself
mycmd="~/Downloads/$rosettapath-
3.5/rosetta_source/bin/relax.linuxgccrelease -database ~/Downloads/$rosettapath-
3.5/rosetta_database/ -s $input -constant_seed -jran $seed -nstruct $nstruct -
out:file:scorefile $f/scores/$nround-$type/score.$c.sc -mute core.io.database -no_filters -
out:overwrite -out:path:pdb $f/output_pdb/$nround-$type -out:suffix .$c > $f/logs/$nround-
$type/run.$c.log" # concaténation de la commande
    echo $mycmd >> "$f/input/$nround-$type/$o" # écrit la
commande

#rmsd
inputname=$(basename $input)

```

```

                                rmsdcmd="--/Downloads/rosettamsterdam-
3.5/rosetta_source/bin/score.linuxgccrelease -database ~/Downloads/rosettamsterdam-
3.5/rosetta_database/ -s $f/output_pdb/$nround-$type/${inputname}.pdb/.$(echo $c)*.pdb} -
native $input -score:empty -out:file:scorefile $f/scores/$nround-$type/rmsd.$c.sc -jran
$seed >> $f/logs/$nround-$type/rmsd.$c.log" # concaténation de la commande
                                echo $rmsdcmd >> "$f/input/$nround-$type/rmsd.$o" # ecrit la
commande
                                remainingjobs=$((remainingjobs-$max))
                                c=$((c+1))
                                done
                                done
                                #
                                # Start runs
                                if [ $mock -eq 0 ]; then sh -c "cat $f/input/$nround-$type/$o |
~/Downloads/parallel-20130622/src/parallel -j $simult_proc -k"; fi #RUN: lancer si mock = 0
                                exitcode=$?
                                if [ $exitcode -gt 0 ]
                                then
                                        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tAn error occured!
Parallel exit status: $exitcode"
                                else
                                        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tEverything is fine,
apparently! Parallel exit status: $exitcode"
                                fi

                                # calcul rmsd
                                echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tCalculating rmsd..."
                                if [ $mock -eq 0 ]; then sh -c "cat $f/input/$nround-$type/rmsd.$o |
~/Downloads/parallel-20130622/src/parallel -j $simult_proc -k"; fi #RUN: lancer si mock = 0
                                exitcode=$?
                                if [ $exitcode -gt 0 ]
                                then
                                        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tAn error occured!
Parallel exit status: $exitcode"
                                else
                                        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tEverything is fine,
apparently! Parallel exit status: $exitcode"
                                fi
                                echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tRmsd calculated!"

                                # Tri scores relax (par total_score ascendant)
                                ## Scores totaux
                                echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tFiltering scores..."
                                head -n 2 $f/scores/$nround-$type/score.$((1+$initial_offset)).sc >
$f/scores/$nround-$type/scores_total.tmp
                                cat $f/scores/$nround-$type/score.*.sc | egrep -v
'(SEQUENCE|description)' >> $f/scores/$nround-$type/scores_total.tmp
                                awk '!x[$0]++' $f/scores/$nround-$type/scores_total.tmp >
$f/scores/$nround-$type/scores_total.sc
                                ## Tri rmsd
                                echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tFiltering rmsd..."
                                head -n 1 $f/scores/$nround-$type/rmsd.$((1+$initial_offset)).sc >
$f/scores/$nround-$type/rmsd_total.tmp
                                cat $f/scores/$nround-$type/rmsd.*.sc | grep -v 'description' >>
$f/scores/$nround-$type/rmsd_total.tmp
                                awk '!x[$0]++' $f/scores/$nround-$type/rmsd_total.tmp >
$f/scores/$nround-$type/rmsd_total.sc
                                ## Tri pour jointure rms scores
                                echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tAssociating scores and
rmsd..."
                                nbscores=$(((-2+(wc -l $f/scores/$nround-$type/scores_total.sc | cut
-d' ' -f 1)))

```



```

        nbrmsd=$((-1+(wc -l $f/scores/$nround-$type/rmsd_total.sc | cut -d'
' -f 1)))
        if [ $nbscores -eq $nbrmsd ]
        then
            TAB=`echo -e "\t"`
            sed 's/ \+/\t/g' $f/scores/$nround-$type/scores_total.sc >
$f/scores/$nround-$type/scores_total.tab #Format tab
            sed 's/ \+/\t/g' $f/scores/$nround-$type/rmsd_total.sc >
$f/scores/$nround-$type/rmsd_total.tab #Format tab
            tail -n +3 $f/scores/$nround-$type/scores_total.tab | sort -
t"$TAB" -k 22 > $f/scores/$nround-$type/scores_total.tmp #Sort sur description
            tail -n +2 $f/scores/$nround-$type/rmsd_total.tab | sort -
t"$TAB" -k 14 > $f/scores/$nround-$type/rmsd_total.tmp #Sort sur description
            cut -f 13 $f/scores/$nround-$type/rmsd_total.tmp | paste -d
"$TAB" $f/scores/$nround-$type/scores_total.tmp - > $f/scores/$nround-$type/scores_rmsd.tmp
#Extraction rms et ajout sur scores
            echo -e "$(head -2 $f/scores/$nround-$type/scores_total.tab
| tail -1)rms" > $f/scores/$nround-$type/scores_total.sc_topall #Entête avec rmsd
            sort -t"$TAB" -k 2,2 -n $f/scores/$nround-
$type/scores_rmsd.tmp >> $f/scores/$nround-$type/scores_total.sc_topall #Sort sur
total_score

        else
            echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tUnmatching scores
and rmsd entries!"
        fi
        ## Filtre rms
        head -1 $f/scores/$nround-$type/scores_total.sc_topall >
$f/scores/$nround-$type/scores_total.sc_top_rms_$(echo "$rmsmin")_$(echo "$rmsmax")
        awk -v min=$rmsmin -v max=$rmsmax '$23 >= min && $23 <= max {print
$0;}' $f/scores/$nround-$type/scores_total.sc_topall >> $f/scores/$nround-
$type/scores_total.sc_top_rms_$(echo "$rmsmin")_$(echo "$rmsmax") #Tri sur rms
        nbitem=$(tail -n +2 $f/scores/$nround-
$type/scores_total.sc_top_rms_$(echo "$rmsmin")_$(echo "$rmsmax") | wc -l)

        ## Nettoyage
        rm $f/scores/$nround-$type/scores_total.tmp
        rm $f/scores/$nround-$type/rmsd_total.tmp
        rm $f/scores/$nround-$type/scores_total.tab
        rm $f/scores/$nround-$type/rmsd_total.tab
        rm $f/scores/$nround-$type/scores_total.sc
        rm $f/scores/$nround-$type/scores_rmsd.tmp
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tFiltering done!"

        # Extraction des top poses
        topsum="$f/scores/$nround-$type/top_poses.sum"
        if [ $nbitem -gt 0 ] #Si au moins un decoy satisfaisant
        then
            echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tGetting top $top
poses from $nbitem decoys with rms between $rmsmin and $rmsmax..."
            echo -e 'total_score\tPose\tPath\trms' > $topsum
            IFS=$'\n'
            for topline in $(echo -e "$(head -n $((top+1))
$f/scores/$nround-$type/scores_total.sc_top_rms_$(echo "$rmsmin")_$(echo "$rmsmax") | tail -
n +2)")
            do
                echo "$topline" | awk -v f="$f/output_pdb/$nround-
$type/" '{print $2"\t"$22"\t"f$22".pdb\t"$23}' >> $topsum #ajout des i_sc / nom / chemin
complet des top poses
            done
            echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tTop poses found!
List summarized in $topsum"
        else #Si aucun decoy dans l'intervalle

```

```

                                echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tUnable to find $top
poses. No decoys with total_score better than input !"
                                fi

                                # Calcul cavités des top poses
                                ## Fichiers .queue
                                echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tGenerating files for
cavities in $top top poses..."
                                if [ $mock -eq 0 ]; then ./pdbsurf-batch.sh "tail -n +2 $topsum | cut
-f3" $simult_proc $max > $f/logs/$nround-$type/pdbsurf-batch.log; fi #RUN: lancer si mock =
0
                                exitcode=$?
                                if [ $exitcode -gt 0 ]
                                then
                                        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tAn error occured!
Pdbsurf-batch exit status: $exitcode"
                                else
                                        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tEverything is fine,
apparently! Pdbsurf-batch exit status: $exitcode"
                                fi
                                ## Run surfracer et 3v
                                echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tSearching for cavities in
$top top poses..."
                                if [ $mock -eq 0 ]; then cat $f/tmp/*.queue | ~/Downloads/parallel-
20130622/src/parallel -j $simult_proc -k > $f/logs/$nround-$type/pdbsurf-auto.log; fi #RUN:
lancer si mock = 0
                                exitcode=$?
                                if [ $exitcode -gt 0 ]
                                then
                                        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tAn error occured!
Parallel exit status: $exitcode"
                                else
                                        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tEverything is fine,
apparently! Parallel exit status: $exitcode"
                                fi
                                #cat $f/tmp/*/pdbsurf-cav.sum | while read line
                                echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tMerging scores from top
poses..."
                                cat $topsum | while read line
                                do
                                        linepdb=$(echo $line | cut -f2)
                                        if [ $linepdb == 'Pose' ]
                                        then
                                                echo -e $line'\t'$(head -n 1 $f/tmp/*/pdbsurf-
cav.sum | cut -f2-) > $topsum.tmp #concaténation lignes 1
                                        else
                                                echo -e $line'\t'$(cat $f/tmp/*/pdbsurf-cav.sum |
grep $linepdb | cut -f2-) >> $topsum.tmp #concaténation des scores (correspondance cherchée
par grep = pas optimal sur de nombreuses lignes)
                                        fi
                                done
                                mv $topsum.tmp $topsum # remplacement du fichier topsum avec ajout
scores
                                echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tScores merged! List
summarized in $topsum"
                                ## Nettoyage
                                rm -R $f/tmp/

;;
# Fin relax
# Docking
'docking') echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tStarting docking"
c=$((1+$initial_offset)) # compteur
if [ $spin -eq 1 ]; then spinoption=' -spin'; else spinoption=''; fi;

```

```

fi;
    if [ $rand1 -eq 1 ]; then rand1option=' -randomize1'; else rand1option='';
fi;
    if [ $rand2 -eq 1 ]; then rand2option=' -randomize2'; else rand2option='';
fi;
    if [ "$unitrans" == "0" ]; then unitransoption=''; else unitransoption=" -
uniform_trans=$unitrans"; fi;
    # Liste inputs
    echo -e "$inputs" | while read i
    do
        input=$(echo -e "$i" | tr -s ' ' | cut -f3) # récupère chaque chemin
complet de fichier

        # Boucle tant qu'il y a des decoys à faire
        remainingjobs=$tot
        while [ $remainingjobs -gt 0 ]
        do
            seed=$((1111111+$c)) #random_seed pas random
            # nombre de structures à faire
            if [ $remainingjobs -lt $max ]; then nstruct=$remainingjobs;
else nstruct=$max; fi

            #
            mycmd="~/Downloads/$rosettopath-
3.5/rosetta_source/bin/docking_protocol.linuxgccrelease -database ~/Downloads/$rosettopath-
3.5/rosetta_database/ -s $input -constant_seed -jran $seed -dock_pert $pert_deg
$pert_dist$spinooption$unitransoption -nstruct $nstruct -out:file:scorefile
$f/scores/$nround-$type/score.$c.sc -mute core.io.database -no_filters -
out:overwrite$rand1option$rand2option -ex1 -ex2aro -out:path:pdb $f/output_pdb/$nround-$type
-out:suffix .$c -docking:partners B_A > $f/logs/$nround-$type/run.$c.log" # concaténation de
la commande #TODO ajuster les paramètres (notamment spin, degres etc.)
            echo $mycmd >> "$f/input/$nround-$type/$o" # ecrit la
commande

            remainingjobs=$((remainingjobs-$max))
            c=$((c+1))

        done
    done
    #
    # Start runs
    if [ $mock -eq 0 ]; then sh -c "cat $f/input/$nround-$type/$o |
~/Downloads/parallel-20130622/src/parallel -j $simult_proc -k"; fi #RUN: lancer si mock = 0
    exitcode=$?
    if [ $exitcode -gt 0 ]
    then
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tAn error occured!
Parallel exit status: $exitcode"
    else
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tEverything is fine,
apparently! Parallel exit status: $exitcode"
    fi

    # Tris scores docking
    echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tFiltering scores..."
    cat $f/scores/$nround-$type/score.$((1+$initial_offset)).sc | grep -v
"SCORE" > $f/scores/$nround-$type/scores_total.tmp
    cat $f/scores/$nround-$type/score.$((1+$initial_offset)).sc | grep
"total_score" >> $f/scores/$nround-$type/scores_total.tmp
    #cat $f/scores/$nround-$type/score.* | grep ${input/.pdb/} >>
$f/scores/$nround-$type/scores_total.tmp #BUG si input basculé d'un tour à l'autre
    cat $f/scores/$nround-$type/score.*.sc | egrep -v
'SEQUENCE|total_score' >> $f/scores/$nround-$type/scores_total.tmp #BUG CORRECTION ne
vérifie plus le nom dans la recherche des decoys mais exclue les entêtes à la place
    awk '!x[$0]++' $f/scores/$nround-$type/scores_total.tmp >
$f/scores/$nround-$type/scores_total.sc
    rm $f/scores/$nround-$type/scores_total.tmp

```

```

TAB=`echo -e "\t"`
sed 's/ \+/\t/g' $f/scores/$nround-$type/scores_total.sc >
$f/scores/$nround-$type/scores_total.tab
head -2 $f/scores/$nround-$type/scores_total.tab | tail -1 >
$f/scores/$nround-$type/scores_total.sc_topall
tail -n +3 $f/scores/$nround-$type/scores_total.tab | sort -t"$TAB" -
n -k 5,5 >> $f/scores/$nround-$type/scores_total.sc_topall #sort sur i_sc
head -2 $f/scores/$nround-$type/scores_total.tab | tail -1 >
$f/scores/$nround-$type/scores_total.sc_top_Isc_-10_-5
awk -v J=5 -v R="^-[5-9].)|(10.000))" '{if (match($J, R)) print
$0;}' $f/scores/$nround-$type/scores_total.sc_topall >> $f/scores/$nround-
$type/scores_total.sc_top_Isc_-10_-5
#head -201 $f/scores/$nround-$type/scores_total.sc_top_Isc_-10_-5 >
$f/scores/$nround-$type/scores_total.sc_top200_Isc_-10_-5
nbitem=$(cat $f/scores/$nround-$type/scores_total.sc_top_Isc_-10_-5 |
wc -l)
mv $f/scores/$nround-$type/scores_total.sc_top_Isc_-10_-5
$f/scores/$nround-$type/scores_total.sc_top`expr $nbitem - 1`_Isc_-10_-5
rm $f/scores/$nround-$type/scores_total.tab
rm $f/scores/$nround-$type/scores_total.sc
echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")] \tFiltering done!"

# Extraction des top poses
topsum="$f/scores/$nround-$type/top_poses.sum"
if [ $nbitem -gt 1 ] #Si au moins un decoy satisfaisant
then
    echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")] \tGetting top $top
poses from $((($nbitem-1)) decoys with i_sc between -10 and -5..."
    echo -e 'I_sc\tPose\tPath\ttrms' > $topsum
    IFS=$'\n'
    for topline in $(echo -e "$(head -n $((($top+1))
$f/scores/$nround-$type/scores_total.sc_top`expr $nbitem - 1`_Isc_-10_-5 | tail -n +2))"
do
        echo "$topline" | awk -v f="$f/output_pdb/$nround-
$type/" '{print $5"\t"$27"\t"f$27".pdb\t"$3}' >> $topsum #ajout des i_sc / nom / chemin
complet des top poses
        done
    echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")] \tTop poses found and
summarized in $topsum"
    else #Si aucun decoy dans l'intervalle
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")] \tUnable to find $top
poses. No decoys with i_sc between -10 and -5 !"
    fi

# Calcul cavités des top poses
## Fichiers .queue
echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")] \tGenerating files for
cavities in $top top poses..."
if [ $mock -eq 0 ]; then ./pdsurf-batch.sh "tail -n +2 $topsum | cut
-f3" $simult_proc $max > $f/logs/$nround-$type/pdsurf-batch.log; fi #RUN: lancer si mock =
0
exitcode=$?
if [ $exitcode -gt 0 ]
then
    echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")] \tAn error occured!
Pdsurf-batch exit status: $exitcode"
else
    echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")] \tEverything is fine,
apparently! Pdsurf-batch exit status: $exitcode"
fi
## Run surfracer et 3v
echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")] \tSearching for cavities in
$top top poses..."

```

```

        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tGenerating files for
cavities in $top top poses..."
        if [ $mock -eq 0 ]; then cat $f/tmp/*.queue | ~/Downloads/parallel-
20130622/src/parallel -j $simult_proc -k > $f/logs/$nround-$type/pdbsurf-auto.log ; fi #RUN:
lancer si mock = 0
        exitcode=$?
        if [ $exitcode -gt 0 ]
        then
            echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tAn error occurred!
Parallel exit status: $exitcode"
        else
            echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tEverything is fine,
apparently! Parallel exit status: $exitcode"
        fi
        #cat $f/tmp/*/pdbsurf-cav.sum | while read line
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tMerging scores from top
poses..."
        cat $topsum | while read line
        do
            linepdb=$(echo $line | cut -f2)
            if [ $linepdb == 'Pose' ]
            then
                echo -e $line'\t'$(head -n 1 $f/tmp/*/pdbsurf-
cav.sum | cut -f2-) > $topsum.tmp #concaténation lignes 1
            else
                echo -e $line'\t'$(cat $f/tmp/*/pdbsurf-cav.sum |
grep $linepdb | cut -f2-) >> $topsum.tmp #concaténation des scores (correspondance cherchée
par grep = pas optimal sur de nombreuses lignes)
            fi
        done
        mv $topsum.tmp $topsum # remplacement du fichier topsum avec ajout
scores
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tScores merged! List
summarized in $topsum"
        ## Nettoyage
        rm -R $f/tmp/

;;
# Fin docking
# Design
'design') echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tStarting design"
#
c=$((1+$initial_offset)) # compteur
# Liste inputs
echo -e "$inputs" | while read i
do
    input=$(echo -e "$i" | tr -s ' ' | cut -f3) # récupère chaque chemin
complet de fichier

# Boucle tant qu'il y a des decoys à faire
remainingjobs=$tot
while [ $remainingjobs -gt 0 ]
do
    seed=$((1111111+$c)) #random_seed pas random
    # nombre de structures à faire
    if [ $remainingjobs -lt $max ]; then nstruct=$remainingjobs;
else nstruct=$max; fi
    #
    mycmd="~/Downloads/$rosettopath-
3.5/rosetta_source/bin/fixbb.linuxgccrelease -database ~/Downloads/$rosettopath-
3.5/rosetta_database/ -s $input -constant_seed -jran $seed -nstruct $nstruct -resfile
$f/resfile-Y -ex1 -ex2 -out:overwrite -out:file:scorefile $f/scores/$nround-

```

```

$type/score.$c.sc -out:path:pdb $f/output_pdb/$nround-$type -out:overwrite -out:suffix .$c >
$f/logs/$nround-$type/run.$c.log" #TODO: customisation resfile
        echo $mycmd >> "$f/input/$nround-$type/$o" # ecrit la
commande
        remainingjobs=$((($remainingjobs-$max))
        c=$((c+1))
done
done
#
#
        # Start runs
        if [ $mock -eq 0 ]; then sh -c "cat $f/input/$nround-$type/$o |
~/Downloads/parallel-20130622/src/parallel -j $simult_proc -k" ; fi #RUN: lancer si mock = 0
        exitcode=$?
        if [ $exitcode -gt 0 ]
        then
                echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tAn error occured!
Parallel exit status: $exitcode"
        else
                echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tEverything is fine,
apparently! Parallel exit status: $exitcode"
        fi

        # identification séquence
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tGetting sequences..."
        echo -e "PDB\tSeq" > $f/scores/$nround-$type/seq_total.tab
        cat $f/scores/$nround-$type/score.*.sc | egrep -v
'(SEQUENCE|total_score)' | while read line
do
        #linepdb=$(echo $line | cut -d' ' -f22) #BUG en cas de
nombre d'espaces changés
        linepdb=$(echo $line | tr -s ' ' | cut -d' ' -f22) #BUG
CORRECTION séparation ok malgré multiples espaces grace à "tr -s ' ' " éliminant les
successions d'espaces
        echo -e "$linepdb\t$(sh -c "./pdb2fasta.sh
$f/output_pdb/$nround-$type/$linepdb.pdb 'A' " | grep -v '>') >> $f/scores/$nround-
$type/seq_total.tab # Recuperation de la sequence de NF (chaîne A)
done
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tSequences extracted (or
not)!"

        # Tri scores relax (par total_score ascendant)
        ## Scores totaux
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tFiltering scores..."
        head -n 2 $f/scores/$nround-$type/score.$((1+$initial_offset)).sc >
$f/scores/$nround-$type/scores_total.tmp
        cat $f/scores/$nround-$type/score.*.sc | egrep -v
'(SEQUENCE|description)' >> $f/scores/$nround-$type/scores_total.tmp
        awk '!x[$0]++' $f/scores/$nround-$type/scores_total.tmp >
$f/scores/$nround-$type/scores_total.sc
        ## Tri pour jointure seq scores
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tAssociating scores and
sequences..."
        nbscores=$((-2+$wc -l $f/scores/$nround-$type/scores_total.sc | cut
-d' ' -f 1)))
        nbseq=$((-1+$wc -l $f/scores/$nround-$type/seq_total.tab | cut -d' '
-f 1)))
        #echo "$nbscores scores pour $nbrmsd rms et $nbseq séquences"
        if [ $nbscores -eq $nbseq ]
        then
                TAB=`echo -e "\t`

```

```

        sed 's/ \+/\t/g' $f/scores/$nround-$type/scores_total.sc >
$f/scores/$nround-$type/scores_total.tab #Format tab
        tail -n +3 $f/scores/$nround-$type/scores_total.tab | sort -
t"$TAB" -k 22 > $f/scores/$nround-$type/scores_total.tmp #Sort sur description
        tail -n +2 $f/scores/$nround-$type/seq_total.tab | sort -
t"$TAB" -k 1 > $f/scores/$nround-$type/seq_total.tmp #Sort sur description
        cut -f 2 $f/scores/$nround-$type/seq_total.tmp | paste -d
"$TAB" $f/scores/$nround-$type/scores_total.tmp - > $f/scores/$nround-$type/scores_seq.tmp
#Extraction seq et ajout sur scores
        echo -e "$(head -2 $f/scores/$nround-$type/scores_total.tab
| tail -1)seq" > $f/scores/$nround-$type/scores_total.sc_topall #Entête avec seq
        sort -t"$TAB" -k 2,2 -n $f/scores/$nround-
$type/scores_seq.tmp >> $f/scores/$nround-$type/scores_total.sc_topall #Sort sur total_score
        else
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tUnmatching scores
and sequences entries!"
    fi

    ## Nettoyage
    rm $f/scores/$nround-$type/scores_total.tmp
    rm $f/scores/$nround-$type/scores_total.tab
    rm $f/scores/$nround-$type/scores_total.sc
    rm $f/scores/$nround-$type/seq_total.tmp
    rm $f/scores/$nround-$type/seq_total.tab
    rm $f/scores/$nround-$type/scores_seq.tmp
    echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tFiltering done!"

    ## Filtre seq
    echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tGetting top $top poses from
unique sequences..."

    # Extraction des top poses
    topsum="$f/scores/$nround-$type/top_poses.sum"
    echo -e
"$Seq\toccurrences\ttotal_score_min\ttotal_score_max\ttotal_score_mean\ttotal_score_sd" >
$f/scores/$nround-$type/top_seq.sum
    echo -e "total_score\tPose\tPath\tSeq" > $topsum.tmp
    tail -n +2 $f/scores/$nround-$type/scores_total.sc_topall | awk -
F"$TAB" '![_[$23]++]' | head -n $top >> $topsum.tmp #récupération des séquences uniques (print
implicite)
        tail -n +2 $topsum.tmp | awk '{print $23}' | while read seq #
récupère les séquences uniques sur le champ 23
        do
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\t\tFocus on
Silicofitin $seq !"
                seqscores=$(tail -n +2 $f/scores/$nround-
$type/scores_total.sc_topall | grep "$seq" | cut -d"$TAB" -f2 | Rscript -e 'd<-scan("stdin",
quiet=TRUE)' -e 'cat(length(d), min(d), max(d), mean(d), sd(d), sep="\t")')
                echo -e "$seq\t$seqscores" >> $f/scores/$nround-
$type/top_seq.sum #Utilisation de R pour moyenne, etc.
        done
        head -n 1 $topsum.tmp > $topsum
        IFS=$'\n'
        for topline in $(tail -n +2 $topsum.tmp)
        do
                echo "$topline" | awk -v f="$f/output_pdb/$nround-$type/"
'${print $2"\t"$22"\t"f$22".pdb\t"$23}' >> $topsum #ajout des total_score / nom / chemin
complet des top poses
        done
        mv $topsum $topsum.tmp
        cat $f/scores/$nround-$type/top_seq.sum | cut -f2- | paste
$topsum.tmp - > $topsum
        echo -e "[$(date +%Y-%m-%d:%H:%M:%S)]\tTop poses found! List
summarized in $topsum"

```

```

        # Calcul cavités des top poses
        ## Fichiers .queue
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tGenerating files for
cavities in $top top poses..."
        if [ $mock -eq 0 ]; then ./pdsurf-batch.sh "tail -n +2 $topsum | cut
-f3" $simult_proc $max > $f/logs/$nround-$type/pdsurf-batch.log ; fi #RUN: lancer si mock =
0
        exitcode=$?
        if [ $exitcode -gt 0 ]
        then
            echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tAn error occured!
Pdsurf-batch exit status: $exitcode"
        else
            echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tEverything is fine,
apparently! Pdsurf-batch exit status: $exitcode"
        fi
        ## Run surfracer et 3v
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tSearching for cavities in
$top top poses..."
        if [ $mock -eq 0 ]; then cat $f/tmp/*.queue | ~/Downloads/parallel-
20130622/src/parallel -j $simult_proc -k > $f/logs/$nround-$type/pdsurf-auto.log ; fi #RUN:
lancer si mock = 0
        exitcode=$?
        if [ $exitcode -gt 0 ]
        then
            echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tAn error occured!
Parallel exit status: $exitcode"
        else
            echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tEverything is fine,
apparently! Parallel exit status: $exitcode"
        fi
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tMerging scores from top
poses..."
        cat $topsum | while read line
        do
            linepdb=$(echo $line | cut -f2)
            if [ $linepdb == 'Pose' ]
            then
                echo -e $line'\t'$(head -n 1 $f/tmp/*/pdsurf-
cav.sum | cut -f2-) > $topsum.tmp #concaténation lignes 1
            else
                echo -e $line'\t'$(cat $f/tmp/*/pdsurf-cav.sum |
grep $linepdb | cut -f2-) >> $topsum.tmp #concaténation des scores (correspondance cherchée
par grep = pas optimal sur de nombreuses lignes)
            fi
        done
        mv $topsum.tmp $topsum # remplacement du fichier topsum avec ajout
scores
        echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tScores merged! List
summarized in $topsum"
        ## Nettoyage
        rm -R $f/tmp/
    ;;
    # Fin design
    # Autres
    *) echo -e "[$(date +"%Y-%m-%d:%H:%M:%S")]\tRound type not found"
    ;;
    esac
done # fin pour chaque round

```


Annexe 1.10. pdbsurf-auto.sh (v1.5)

```
#!/bin/sh
#=====
# Script: pdbsurf-auto.sh
# Auteur: Simon Huet, 2015, contact@simon-huet.fr
#-----
#ChnG v1.0 : Version initiale basée sur la v1.15 de Pdbsurf.sh (YH Sanejouand, Avril 2013)
#ChnG v1.1 : Recherche de Surface Racer dans ~/Test/scripts/surfaceracer/ en priorité
#ChnG v1.2 : Compare un PDB en complexe AB et les PDB de A et B
#ChnG v1.3 : Ajout de 3v pour calcul de volume de cavité à l'interface
#ChnG v1.4 : Intégration au système de batch docking/design/relax
#ChnG v1.5 : Tentative de considération cavité par cavité pour volumes via 3v (plus lent,
plus précis?)
version='v1.5, April 2015.'

#Name of the executable:
racerexe=surfrace5_0_linux_64bit

#Name of the file with the radii:
rfile=radii.txt

if [ $# -gt 0 ]
then
  nfil=`ls -1d $* | wc -l`
else
  nfil=0
fi
if [ $nfil -eq 0 ]
then
  echo 'Usage: pdbsurf "pdb-file-selection"'
  echo 'Action: Accessible surface area for a set of PDB files, using Surface Racer.'
  exit
fi
prognm='.Pdbsurf>'
echo $prognm $version

echo $prognm 'Accessible surface area for a set of' $nfil 'PDB file(s).'
progwn='%Pdbsurf-Wn>'
proger='%Pdbsurf-Er>'

#CoM Surface Racer and radii file are expected:
#CoM first in local directory; then in HOME directory.
#CoM v1.1: preferred location: /comptes/E072042G/Test/scripts/surfaceracer

#La chaine ne doit pas etre vide mais sait-on jamais ?
if [ -x "$racerexe" -a -n "$racerexe" ]
then
  racesurf=$racerexe
else
  racesurf=`find ~/Downloads/surface_racer_5.0_64bit/ -name $racerexe | tail -1`
  if [ ! -x "$racesurf" -o ! -n "$racesurf" ]
  then
    racesurf=`find . -name $racerexe | tail -1`
    if [ ! -x "$racesurf" -o ! -n "$racesurf" ]
    then
      racesurf=`find ~ -name $racerexe | tail -1`
      if [ ! -x "$racesurf" -o ! -n "$racesurf" ]
      then
        # Not perfect: the database for locate could need an update; slow otherwise.
        racesurf=`locate $racerexe | tail -1`
        if [ ! -x "$racesurf" -o ! -n "$racesurf" ]
        then
```

```

                                echo $proger 'Surface Racer executable:' $racerexe 'not
found.'
                                exit
                                fi
                                fi
                                fi
                                cp $racesurf .
fi
echo $progrnm 'Executable to be used locally:' $racesurf

if [ -f $rfile ]
then
    echo $progrnm 'Local radii file to be used:' $rfile
else
    racedir=`dirname $racesurf`
    rfilnam=$racedir/'$rfile'
    if [ -f $rfilnam ]
    then
        echo $progrnm 'Radii file:' $rfilnam 'copied locally.'
        cp $rfilnam .
    else
        echo $proger 'Radii file:' $rfile 'not found in directory' $racedir
        exit
    fi
fi

#echo ' Directory where to save output PDB file(s) ? (or None)'
#read savdir
savdir='output'
tstwd=`echo $savdir | tr '[a-z]' '[A-Z]`
echo "savedir test: $tstwd"
case $tstwd in
NONE|NON|NO|N) savdir=NONE ;;
*)
if [ ! -d $savdir ] ; then mkdir $savdir ; fi
echo $progrnm 'Output PDB file(s) to be saved in directory:' $savdir ;;
esac

#CoM Surface Racer has sometimes problems, the symptom being a very low SA.
if [ ! $savdir = NONE ]
then
    echo ' Minimum accessible surface expected for saving results ? (<0: Any)'
    #read minsrf
    minsrf=0
    case $minsrf in *(!0-9,!\\.)* ) minsrf=-1 ;; esac
fi

# Input for Surface Racer:
# -----
# Radii set (1 or 2)
# PDB
# Probe (water) radius (1.4: valeur du papier)
# SA (1) SA+MSA (2) or SA+MSA+Curvature (3)
cat > surface_racer.inp_pdbsurf << EOFsr
1
structure.pdb
1.4
1
EOFsr

if test -f pbsurf.sum
then

```

```

echo $progwn 'pdbsurf.sum file already exists. A copy is saved.'
mv pdbsurf.sum pdbsurf.sum_old
fi
echo $prognm $version > pdbsurf.sum
echo $prognm $version > pdbsurf-cav.sum
echo -e "PDB\tInt_surf\tPolar_int_surf\tNonpolar_int_surf\tInt_cavity\tInt_cav_volume" >
pdbsurf-cav.sum

# =====
# Input files are scanned:
# =====
nlowsa=0
nfnnd=0
npb=0
for j in `ls -d $*` # all complex PDB
do
    # counters
    intcav=0
    intsurf=0
    intpolsurf=0
    intnpolsurf=0
    iloop=0

    for i in "$j" "$j.cA" "$j.cB" # for AB, A and B
    do
        iloop=$((iloop+1))
        if [ -f $i ] ; then
            echo $prognm 'File considered:' $i

            # Gzip eventual:
            if [ ${i##*.} = "gz" ]
            then
                zcat $i > structure.pdb
            else
                cp $i structure.pdb
            fi

            natom=`cut -c1-4 structure.pdb | egrep -c 'ATOM|HETA'`
            if [ $natom -le 0 ]
            then
                echo $progwn 'No line with ATOM or HETA keys, as expected in a PDB file.'
            else
                # =====
                # Surface Racer:
                # =====
                rm -f structure.txt structure_residue.txt result.txt pdbsurf.out
                ./$racerexe < surface_racer.inp_pdbsurf > pdbsurf.out

                if [ -f structure.txt -a -f pdbsurf.out ] ; then
                    nok=`grep -A1 'Solvent accessible surface areas' pdbsurf.out | grep -c
'='`
                else
                    nok=0
                fi

                # Calculation ended normally:
                if [ $nok -gt 0 ]
                then
                    srout=`grep -A1 'Solvent accessible surface areas' pdbsurf.out | tail -1`

                # Nombre de cavites trouvees:
                ncav=`cut -c1-6 structure.txt | grep -c CAVITY`

```

```

        echo $prognm $ncav 'cavities found.'

#       Nombre de CA:
        nca=`grep 'ATOM' structure.txt | cut -c 12-17 | grep -c ' CA '`
        echo $i $natom 'atoms,' $nca 'aa,' $ncav 'cavities,' $srout | sed 's/
./,/g' >> pdbsurf.sum

# Counting interface values
if [ $iloop -eq 1 ]
then
        intcav=$ncav
        intsurf=$(echo 0 - $(echo "$srout" | sed -n "s/.*Total area = \([0-9]*\.[0-9]*\).*\1/p") | bc)
        intpolsurf=$(echo 0 - $(echo "$srout" | sed -n "s/.*Polar area= \([0-9]*\.[0-9]*\).*\1/p") | bc)
        intnpolsurf=$(echo 0 - $(echo "$srout" | sed -n "s/.*Non-polar area= \([0-9]*\.[0-9]*\).*\1/p") | bc)
        else
        intcav=$((intcav - $ncav))
        intsurf=$(echo $intsurf + $(echo "$srout" | sed -n "s/.*Total area = \([0-9]*\.[0-9]*\).*\1/p") | bc)
        intpolsurf=$(echo $intpolsurf + $(echo "$srout" | sed -n "s/.*Polar area= \([0-9]*\.[0-9]*\).*\1/p") | bc)
        intnpolsurf=$(echo $intnpolsurf + $(echo "$srout" | sed -n "s/.*Non-polar area= \([0-9]*\.[0-9]*\).*\1/p") | bc)
        fi

#       =====
#       Sauvegarde:
#       =====
        if [ ! $savdir = NONE ]
        then
                satot=`echo $srout | cut -f2 -d'=' | cut -f1 -d',' | sed 's/ //g'`
                if [ `echo "$satot > $minsrf" | bc` -eq 1 ]
                then
                        filnam=`basename $i`

#       On extrait l'info sur les cavites:
                if [ $ncav -gt 0 ]
                then
#       Numero de la premiere ligne avec la cle CAVITY:
                        nfrst=`cut -c1-6 structure.txt | grep -n CAVITY | head -1 | cut -f1
-d':`
                        nfrst=`expr $nfrst - 1`
                        nlns=`wc -l structure.txt | cut -f1 -d' '`
                        nlast=`expr $nlns - $nfrst`

                        head -$nfrst structure.txt > structure_nocav.txt
                        tail -$nlast structure.txt > structure_cav.txt
                        mv structure_nocav.txt structure.txt
                        mv structure_cav.txt $savdir/$filnam'_surfcav'
#       Si pas de cavité, tout de même créer un fichier vide
                        else
                                echo '' > $savdir/$filnam'_surfcav'
                        fi

#       On garde l'entete du fichier PDB:
                        nfrst=`cut -c1-4 structure.pdb | egrep -n 'ATOM|HETA' | head -1 | cut -f1
-d':`
                        nfrst=`expr $nfrst - 1`

#       Output in true PDB format, with ASA instead of B-factors:
#       (Surface Racer gives radius, ASA, MSA, Curvature, after x,y,z)
                        cut -c1-54 structure.txt > structure_xyz.txt

```

```

#      Alignement des virgules:
      cut -c61-68 structure.txt | awk '{ if ( $1 < 10 ) print " " $1 ; else
print $1 }' > structure_surf.txt

      head -$nfrst structure.pdb > structure.txt
      paste -d';' structure_xyz.txt structure_surf.txt | sed 's/;/ /' >>
structure.txt

      echo END >> structure.txt
      mv structure.txt $savdir/$filnam'_surf'

#      Surface minimum respectee:
      else
      nlowsa=`expr $nlowsa + 1`
      fi
#      Sauvegardes demandees:
      fi

      nfnf=`expr $nfnf + 1`
      lastok=$i
      else
      echo $proger 'Surface Racer failed.'
      tail -10 pdbsurf.out
      echo '...'
      echo $proger 'Surface Racer failed for file' $i > pdbsurf.out_lastpb
      echo $proger 'Surface Racer output:' >> pdbsurf.out_lastpb
      cat pdbsurf.out >> pdbsurf.out_lastpb
      npb=`expr $npb + 1`
      lastpb=$i
      fi

# PDB file looks fine:
      fi
# File exists:
      fi
done
echo $progrnm 'Running 3V for cavity volume'
# Recuperation des atomes des cavites uniquement à l'interface
filnam=`basename $j`
grep -f $savdir/'$filnam'.cA_surfcav' -x -v $savdir/'$filnam'_surfcav' | grep -f
$savdir/'$filnam'.cB_surfcav' -x -v > diff.pdb #Ajout option -x pour comparer lignes
entières (et normalement, bien individualiser les cavités)
# Conversion et calcul de volume de cavite
vpath='/comptes/E072042G/Downloads/3v/'
if [ ! -f 'atmtypennumbers' ]
then
      cp $vpath'xyzr/atmtypennumbers' ./
fi
## v.1.4 et inférieur: ensemble des cavités traitées
#$vpath'xyzr/pdb_to_xyzr' ./diff.pdb > ./diff.xyzr
#$vpath'bin/Volume.exe' -i ./diff.xyzr -p 3 -g 0.5 2> ./diff.err > ./diff.txt
#cavol=`cat ./diff.txt | cut -f 3`
#if [ -z $cavol ] ; then cavol=0 ; fi
## v1.5 et supérieur: cavités individuelles
diffstarted=0 # compteur
cavol=0
cat diff.pdb | while read line
do
      debut=$(echo $line | cut -c 1-5)
      if [ "$debut" == "ATOM " ] #Nous sommes dans une cavité
      then
            echo -e "$line" >> diff.tmp.pdb
      else #Nous sommes avant ou après
            if [ $diffstarted -eq 0 ]

```

```

        then # 1ere ligne = pas de cavité encore
            echo '' > diff.tmp.pdb
        else
            $vpath'xyzr/pdb_to_xyzr' ./diff.tmp.pdb > ./diff.tmp.xyzr
            $vpath'bin/Volume.exe' -i ./diff.tmp.xyzr -p 3 -g 0.5 2>
./diff.tmp.err > ./diff.tmp.txt
            cavoltmp=`cat ./diff.tmp.txt | cut -f 3`
            if [ -z $cavoltmp ] ; then cavoltmp=0 ; fi
            echo $progrnm "Cavity volume: $cavoltmp"
            cavol=$(echo $cavol + $cavoltmp | bc)
            echo "$cavol" > diff.tmp.cavol
            echo '' > diff.tmp.pdb
        fi
    fi
    diffstarted=$((diffstarted+1))
done
cavol=$(cat diff.tmp.cavol)
echo $progrnm "Total volume of cavities: "$cavol
##
intsurf=$(echo 'scale=2;' $intsurf/2 | bc)
intpolsurf=$(echo 'scale=2;' $intpolsurf/2 | bc)
intnpolsurf=$(echo 'scale=2;' $intnpolsurf/2 | bc)
echo -e "$j\t$intsurf\t$intpolsurf\t$intnpolsurf\t$intcav\t$cavol" >> pdbsurf-cav.sum

done
if [ ! $savidr = NONE ] ; then cp pdbsurf.sum $savidr/pdbsurf.sum_sav ; fi

case $nfnd in
0) echo $proger 'No successful surface calculation could be performed.' ; exit ;;
1) echo $progrnm 'Surface calculated for file:' $lastok ;;
*) echo $progrnm 'Surface calculated for' $nfnd 'files.' ;;
esac
esac
if [ $nlowsa -gt 0 ] ; then echo $progwn 'Surface lower than' $minsrf 'for' $nlowsa
'file(s).' ; fi

case $npb in
0) echo $progrnm 'Everything went fine, it seems !' ;;
1) echo $progwn 'Surface could not be calculated for file:' $lastpb ;;
*) echo $progwn 'Surface could not be calculated for' $npb 'files.' ;;
esac
esac

#Some cleaning:
rm -f structure.pdb structure_xyz.txt structure_surf.txt structure.txt structure_cav.txt
surface_racer.inp_pdbsurf
rm -f structure_residue.txt result.txt pdbsurf.out
rm -f diff.pdb diff.xyzr diff.txt atmtypenumbers diff_residue.txt diff.surfcav.pdb
diff.txtfcav.pdb diff.err
rm -f diff.tmp.pdb diff.tmp.xyzr diff.tmp.txt diff.tmp.cavol
exit

```

Annexe 1.11. pdb2fasta.sh

```

#!/bin/bash

# 2015 - S. Huet
# 2012 - P. Poulain

usage="Usage: $0 file.pdb [chain]"

#=====
# input data
#=====
# check number of arguments

```

```

if [ ! $# -eq 1 ]
then
    if [ ! $# -eq 2 ]
    then
        echo "Argument error" 1>&2
        echo $usage 1>&2
        exit 1
    fi
fi

# check first argument is an existing regular file
if [ ! -f $1 ]
then
    echo "$1 is not a regular file" 1>&2
    echo $usage 1>&2
    exit 1
fi

name=$1

# check second argument as an optional chain id
if [ -n "$2" ]
then
    ichain=$2
fi

#####
# functions
#####
# list chains in PDB
list_chain() {
    awk '/^ATOM/ && $3 == "CA" {print $5}' | uniq
}

# extract residue sequence from ATOM lines
# take residue from CA atom since
# there is one CA atom per residue
extract_seq_chain() {
    awk -v ch=$1 '/^ATOM/ && $3 == "CA" && $5 == ch {print $4}'
}

# convert newline by space
# for sed lowers this works too: sed ':a;N;$!ba;s/\n/ /g'
remove_newline() {
    tr '\n' ' '
}

# convert 3-letter residue code to 1-letter
convert_aa() {
    sed
    's/ALA/A/g;s/CYS/C/g;s/ASP/D/g;s/GLU/E/g;s/PHE/F/g;s/GLY/G/g;s/HIS/H/g;s/ILE/I/g;s/LYS/K/g;s
    /LEU/L/g;s/MET/M/g;s/ASN/N/g;s/PRO/P/g;s/GLN/Q/g;s/ARG/R/g;s/SER/S/g;s/THR/T/g;s/VAL/V/g;s/T
    RP/W/g;s/TYR/Y/g'
}

# remove space between residues
remove_space() {
    sed 's/ //g'
}

# split fasta sequence at 60 characters (easier to read than 80)
split_60() {
    fold -w 60
}

```

```

}

#=====
# list chain in PDB file
#=====
chains=$(cat $name | list_chain)

#=====
# try to extract a sequence
#=====
for chain in $chains
do
    if [ -n "$ichain" ]
    then
        if [ "$chain" == "$ichain" ]
        then
            sequence=$(cat $name | extract_seq_chain $chain | remove_newline |
convert_aa | remove_space)
            if [[ -n $sequence ]]
            then
                size=$(echo $sequence | wc -c)
                size=$((size-1))
                echo ">${name%.pdb} | chain $chain | $size aa"
                #echo $sequence | split_60
                echo $sequence #no need to split seq lines in 60 chars
            fi
        fi
    else
        sequence=$(cat $name | extract_seq_chain $chain | remove_newline |
convert_aa | remove_space)
        if [[ -n $sequence ]]
        then
            size=$(echo $sequence | wc -c)
            size=$((size-1))
            echo ">${name%.pdb} | chain $chain | $size aa"
            #echo $sequence | split_60
            echo $sequence #no need to split seq lines in 60 chars
        fi
    fi
done

```

Annexe 1.12. pdbsurf-batch.sh

```

#!/bin/sh
#CoM=====
#CoM...pdbsurf-batch.sh: External accessible surface area for a set of proteins.
#CoM-----
#CoM           Internal cavities are also sought for.
#CoM
#CoM Purpose   : Script for batched jobs for Surface Racer.
#CoM
#CoM Arguments : Command line listing PDB filename(s)
#CoM
#CoM Output    : PDB file(s) and Surface Racer job lists.
#CoM-----
#ChnG v1.0
#ChnG v1.1 : Integration dans batchs docking/design/relax
version='v1.1, March 2015.'

#Name of the executable:
racerexe=pdbsurf-auto.sh

```



```

#Number of structures per job
maxjobs=1000

#Number of simultaneous proc
simultproc=1

#Checking input parameter
if [ $# -gt 0 ]
then
  echo 'Starting pbsurf-batch'
  input="$1"
  totjobs=$(eval $input | wc -l)
  if [ $# -gt 1 ] #simultproc changé
  then
    simultproc="$2"
  fi
  if [ $# -gt 2 ] #maxjobs changé
  then
    maxjobs="$3"
  fi
else
  echo 'Usage: pbsurf-batch "command-line-resulting-in-pdb-file-selection"'
  echo 'Action: Accessible surface area for a set of PDB files, using Surface Racer jobs.'
  exit
fi
prognm='.Pbsurf-batch>'
echo $prognm $version
echo $prognm 'Accessible surface area for a set of PDB file(s) listed by: '$input
echo $prognm 'Working directory: '${PWD}
echo $prognm 'Total samples: '$totjobs
echo $prognm 'Max samples per batch: '$maxjobs
echo $prognm 'Simultaneous proc: '$simultproc
prognw='%Pbsurf-Wn>'
proger='%Pbsurf-Er>'

#Random directory function
tmpname(){ cat /dev/urandom | head -c32 | md5sum | cut -f1 -d' '; }
tmpdir=`tmpname`

#Create the tmp folder
if [ ! -d 'tmp' ]
then
  echo $prognm 'Creating tmp directory'
  mkdir 'tmp'
fi

#How many jobs?
nbjobs=$(awk -v tot="$totjobs" -v proc="$simultproc" 'BEGIN { rounded = sprintf("%.0f",
tot/proc); print rounded }')
if [ $nbjobs -gt $maxjobs ]
then
  nbjobs=$maxjobs
fi
echo $prognm 'Number of jobs queued per proc: '$nbjobs

#Get the lists
eval "$input" | awk -v tmp="$tmpdir" -v nb="$maxjobs" '
function cmd( c )
{
  while( (c|getline foo) > 0 )
    printf( "%s\n", foo );
  close( c );
}

```

```

}
BEGIN{
    x="1";
    n="0";
}
{
    print > ("tmp/" tmp "_" x ".list");
    n=n+1;
}
NR%(nb)==0{
    close("tmp/" tmp "_" x ".list");
    mkdir ("tmp/" tmp "_" x);
    x=x+1;
}
END{
    "ls tmp/" tmp "*.list | wc -l" | getline j;
    print n > ("tmp/" tmp ".nsample");
    print j+0 > ("tmp/" tmp ".nbatch");
}
,
#Get the number of samples/batches back
nsample=`cat 'tmp/'$tmpdir'.nsample`
nbatch=`cat 'tmp/'$tmpdir'.nbatch`
echo $prognm 'Number of samples for task #'$tmpdir': '$nsample
echo $prognm 'Number of Surface Racer batches for task #'$tmpdir': '$nbatch

#Loop for each batch
for ((i=1; i<=$nbatch; i++))
do
    #Creating folders
    echo $prognm 'Creating directory: tmp/'$tmpdir'_'$i
    mkdir 'tmp/'$tmpdir'_'$i
    #Copying files
    echo $prognm '... and copying files listed in: tmp/'$tmpdir'_'$i'.list'
#    cat 'tmp/'$tmpdir'_'$i'.list' | xargs -I % cp % 'tmp/'$tmpdir'_'$i
    cat 'tmp/'$tmpdir'_'$i'.list' | xargs -I{} sh -c 'cp "$0" tmp/'$tmpdir'_'$i' && grep
-e "^ATOM[\ ]\{1,\}[0-9]\{1,\}[\ ]\{1,\}[0-9a-zA-Z]\{1,\}[\ ]\{1,\}[a-zA-Z]\{1,\}[\ ]\{1,\}A" "$0" > tmp/'$tmpdir'_'$i'/${basename "$0"}.cA && grep -e "^ATOM[\ ]\{1,\}[0-9]\{1,\}[\ ]\{1,\}[0-9a-zA-Z]\{1,\}[\ ]\{1,\}[a-zA-Z]\{1,\}[\ ]\{1,\}B" "$0" > tmp/'$tmpdir'_'$i'/${basename "$0"}.cB' {}
done

#Generating pdbsurf-auto.sh inputs
echo $prognm 'Generating queues for pdbsurf-auto.sh: tmp/'$tmpdir'.queue'
tmpqueue=`find "${PWD}" -type d -name $tmpdir* | awk -v pwd="$PWD" '{print "cd " $0 " && ../../pdbsurf-auto.sh " $0 "/*.pdb"}'`
echo "$tmpqueue" > 'tmp/'$tmpdir'.queue'
# Adding chains A and B separated in the queue lists
#echo "$tmpqueue" | awk '{gsub(/\ .pdb/, ".pdb.cA"); print $0;} ' >> 'tmp/'$tmpdir'.queue'
#echo "$tmpqueue" | awk '{gsub(/\ .pdb/, ".pdb.cB"); print $0;} ' >> 'tmp/'$tmpdir'.queue'

#case $npb in
#0) echo $prognm 'Everything went fine, it seems !' ;;
#1) echo $prognm 'Surface could not be calculated for file:' $lastpb ;;
#*) echo $prognm 'Surface could not be calculated for' $npb 'files.' ;;
#esac

#Some cleaning:
#rm -f structure.pdb structure_xyz.txt structure_surf.txt structure.txt structure_cav.txt
#rm -f structure_residue.txt result.txt pbsurf.out

#Byebye
exit

```

Annexe 1.13. crontab.sh

```
# Script récursif pour la surveillance d'un batch en cours par email (via crontab)
cd [chemin du répertoire de travail] && echo -e "$(echo -e "Dernière lignes de
$PWD/batch.log : " && tail -n 30 batch.log && echo -e '\n' && top -bn2 | egrep
'^(top|Task|Cpu|Mem|Swap)' | tail -n 5 && echo -e '\nps -u [nom d'utilisateur] r : ' && ps
-u [nom d'utilisateur] r && find scores/ -maxdepth 1 -name '*-*' | while read round; do
echo -e '\n'$round'/score.*.sc lines: '$(cat $round/score.*.sc | wc -l); done)" | mailx -s
"Log du job en cours dans $PWD" [courriel du destinataire à informer]
```

Annexe 2. *Productions scientifiques*

Annexe 2.1. Publication

Huet, S., Gorre, H., Perrocheau, A., Picot, J., & Cinier, M. (2015). *Use of the Nanofitin alternative scaffold as a GFP-ready fusion tag*. *PLoS ONE*, 10(11): e0142304.

Résumé :

With the continuous diversification of recombinant DNA technologies, the possibilities for new tailor-made protein engineering have extended on as an on-going basis. Among these strategies, the use of the green fluorescent protein (GFP) as a fusion domain has been widely adopted for cellular imaging and protein localization. Following the lead of the direct head-to-tail fusion of GFP, we proposed to provide additional features to recombinant proteins by genetic fusion of artificially derived binders. Thus, we reported a GFP-ready fusion tag consisting in a small and robust fusion-friendly anti-GFP Nanofitin binding domain as a proof-of-concept. While limiting steric effects on the carrier, the GFP-ready tag allows the capture of GFP or its blue (BFP), cyan (CFP) and yellow (YFP) alternatives. Here, we described the generation of the GFP-ready tag from the selection of a Nanofitin variant binding to the GFP and its spectral variants with a nanomolar affinity, while displaying a remarkable folding stability, as demonstrated by its full resistance upon thermal sterilization process or the full chemical synthesis of Nanofitins. To illustrate the potential of the Nanofitin-based tag as a fusion partner, we compared the expression level in *Escherichia coli* and activity profile of recombinant human tumor necrosis factor alpha (TNF α) constructs, fused to a SUMO or GFP-ready tag. Very similar expression levels were found with the two fusion technologies. Both domains of the GFP-ready tagged TNF α were proved fully active in ELISA and interferometry binding assays, allowing the simultaneous capture by an anti-TNF α antibody and binding to the GFP, and its spectral mutants. The GFP-ready tag was also shown inert in a L929 cell based assay, demonstrating the potent TNF α mediated apoptosis induction by the GFP-ready tagged TNF α . Eventually, we proposed the GFP-ready tag as a versatile capture and labeling system in addition to expected applications of anti-GFP Nanofitins (as illustrated with previously described state-of-the-art anti-GFP binders applied to living cells and *in vitro* applications). Through a single fusion domain, the GFP-ready tagged proteins benefit from subsequent customization within a wide range of fluorescence spectra upon indirect binding of a chosen GFP variant.

Annexe 2.2. Communications orales

Huet S., Cinier M., Sanejouand Y-H. (2014). **Rational design of Nanofitins targeting a defined epitope.** *Journées Scientifiques de l'Ecole Doctorale Biologie-Santé, 10 Décembre 2014 (15 minutes).*

Résumé :

Nanofitins belongs to the family of affinity protein scaffold, and constitutes as such an alternative to antibodies. They offer the advantages of being small, single chain, extremely stable to both temperature and pH, as well as being highly expressed in simple expression systems like *E. coli*. Their generation using *in vitro* selection system allows to obtain hits from large numbers of variants, which need then to be screened out for the desired properties. This screening effort can be further densified when the targeting of a very defined epitope is required. Assisted-generation of Nanofitins through the computed design of interaction surfaces would be useful to complement this approach and maximize the screening outputs since it would allow to target specific regions of the protein of interest, generally leading to predictable effect of the designed binder.

We describe a computational method for the rational design of Nanofitins binding a targeted surface patch of interest on a macromolecule. Prediction algorithms enforced by published datas are used to identify favorable regions of the target surface for protein-protein interaction. Complexes generated in this way are used as a starting point from which the Nanofitin interaction surface is subjected to *de novo* protein design, and sequences providing the highest complex stability are sorted out. The method was successfully used to perform the entire *in silico* design of Nanofitins binding a selected surface patch of the Green Fluorescent Protein, as confirmed by *in vitro* assay.

This work provides a promising method to generate designed binding proteins that, if combining high affinity for their target and strong specificity of the interaction region, would narrow down the screening effort to generate therapeutically relevant Nanofitins.

Huet S., Cinier M. (2015). **Integration of the OctetRED96 System in the Nanofitin Discovery Platform.** *FortéBio user meeting, 4 Juin 2015 (30 minutes).*

Résumé :

Affilogic is a biotech company developing a new class of affinity ligands, Nanofitins®, which are able to selectively bind many targets and have proven excellent tools for:

- Targeting (immunolocalization, in vivo neutralization): Therapeutic applications
- Capture (affinity chromatography, protein removal): Bioprocessing applications
- Detection (immunoassays, Western Blot): Diagnostic applications

These antibody-mimetics demonstrate superior pharmaceutical properties: high affinity and controlled specificity, excellent druggability (very stable > 65°C and highly soluble), fast generation process and cost-effective manufacturing in standard bacterial strains.

Thanks to these competitive advantages, Nanofitin-based drugs aim at overcoming antibodies limitations, in particular by allowing better tissue penetration and patient-friendly formulations. Affilogic is investigating with several industrial and academic partners therapeutic fields such as Inflammation (FP7 programme), Dermatology (topical treatment for inflammatory skin diseases), CNS (assessing delivery through BBB), Infectious diseases (resistant Gram-negative bacteria through targeting approach).

Annexe 2.3. Posters

Huet S., Cinier M., Sanejouand Y-H. (2014). **Transfert de reconnaissance moléculaire : de la Nanofitine vers l'humain.** *Journées Scientifiques de l'Ecole Doctorale Biologie-Santé, 9-10 Décembre 2014.*

Résumé :

Les anticorps sont aujourd'hui les molécules de référence dans le domaine de la reconnaissance moléculaire. Ils sont utilisés dans de nombreux formats et peuvent ainsi être recombinants ou non, provenir de différents organismes (anticorps humains, murins, camélidés...), être complets (IgG) ou être composés uniquement de fragments d'anticorps (Fab, scFv...). En supplément de ces anticorps et dérivés d'anticorps, l'arsenal des protéines d'affinité est complété par des charpentes alternatives qui se sont développées au cours des dernières décennies. Les Nanofitines (NFs) appartiennent à cette dernière catégorie. Sélectionnées à partir de banques combinatoires de la protéine bactérienne Sac7d, elles combinent les avantages d'une extrême stabilité (à la température et au pH), d'une petite taille (10 kDa, donnant accès à une meilleure pénétration tissulaire), d'une forte expression en *E. coli* (pour un faible coût de production) et d'une flexibilité d'ingénierie.

A ce jour, les anticorps thérapeutiques bénéficiant d'autorisations de mise sur le marché sont essentiellement humains ou humanisés, le but étant de minimiser le risque de réaction immunitaire vis-à-vis de l'anticorps lors du traitement des patients. En pratique, cela signifie le plus souvent que des anticorps issus d'autres organismes (par exemple la souris) sont modifiés de sorte à ressembler au mieux à leur homologue humain. Nous présentons ici une approche similaire d'humanisation des Nanofitines. Dans ces travaux, le site de liaison de NFs est greffé sur de nouvelles charpentes d'origine humaine découvertes par criblage d'une banque de données de structures 3D. La preuve de concept est réalisée à l'aide de NFs fixant spécifiquement la GFP (Green Fluorescent Protein) avec des affinités mesurées entre 1 et 10 nM sur la charpente de Sac7d. La caractérisation des NFs anti-GFP générées *in vitro* par ribosome display ainsi que celle des charpentes humaines après transfert de sites de fixation ont été réalisées à l'aide de mesures par ELISA, interférométrie de couche biologique et modification d'émission de fluorescence.

Huet S. (2014). **Reconnaissance moléculaire: de l'*in silico* à l'*in vitro* et vice versa.** *Les Doctoriales 2014, 23-28 Mars 2014.* Poster présenté lors d'une semaine de rencontres entre entreprises et jeunes chercheurs issus d'écoles doctorales pluridisciplinaires.

Transfert de reconnaissance moléculaire: de la Nanofitine vers une protéine humaine

HUET, Simon^{1,2,*}; CINIÉR, Mathieu²; SANEJOUAND, Yves-Henri¹

* Contact: simon.huet@univ-nantes.fr; ¹ UFIP, FRE CNRS 3478, UFR Sciences et Techniques, 2 rue Houssinière, 44322 Nantes; ² Affilogic SAS, UFR Sciences et Techniques, 2 rue Houssinière, 44322 Nantes

Protéines de reconnaissance moléculaire

Les anticorps, molécules de référence dans le domaine de la **reconnaissance moléculaire**, peuvent provenir de différents organismes (anticorps humains, murins, camélidés...) et existent dans de nombreux formats : anticorps complets (IgG) ou fragments d'anticorps (Fab, scFv...). En supplément de ces **anticorps et dérivés d'anticorps**, l'arsenal des protéines de reconnaissance est complété par une catégorie de **charpentes alternatives**, qui s'est développée récemment et à laquelle appartiennent les **Nanofitines (NFs)**. Sélectionnées pour leur **affinité** et leur **spécificité** envers une cible donnée à partir de banques combinatoires de variants de la protéine bactérienne Sac7d, elles combinent les avantages d'une **extrême stabilité** (à la température et au pH), d'une **petite taille** (10 kDa, pour une meilleure pénétration tissulaire), d'une **forte expression en E. coli** (pour un faible coût de production) et d'une **flexibilité d'ingénierie**.

Molécules thérapeutiques et humanisation

A ce jour, les **anticorps thérapeutiques** bénéficiant d'une autorisation de mise sur le marché sont essentiellement **humains ou humanisés** (issus d'autres organismes, ils sont modifiés de sorte à ressembler au mieux à leur homologue humain), le but étant notamment de **minimiser le risque de réaction immunitaire**. Nous présentons ici une **approche d'humanisation des Nanofitines** par transfert de leurs sites de reconnaissance. Toutefois, les régions non-variables des Nanofitines étant non immunogènes, ce transfert de charpente a pour but **d'accéder à de nouvelles propriétés intrinsèques** aux protéines humaines (comme une extension de la demi-vie).

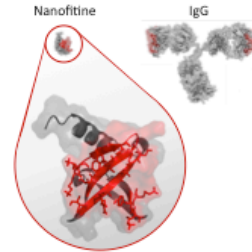


Figure 1: Structure d'une Nanofitine. Les Nanofitines sont en moyenne 15 fois plus petites que des anticorps (IgG). Leurs sites de reconnaissance sont figurés en rouge.

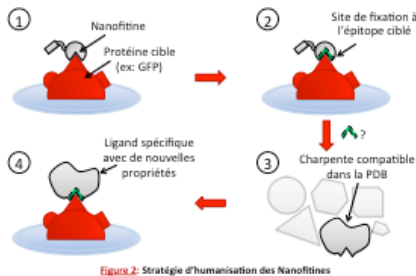


Figure 2: Stratégie d'humanisation des Nanofitines

Stratégie du transfert de site de fixation d'une Nanofitine vers une nouvelle charpente (Fig. 2)

1. Générer *in vitro* des Nanofitines spécifiques d'un épitope d'une cible d'intérêt thérapeutique (démonstration du concept sur la Green Fluorescent Protein ou GFP).
2. Identifier les coordonnées spatiales des acides aminés impliqués dans leur fixation spécifique.
3. Cribler les structures 3D présentes dans la Protein Data Bank (PDB) pour trouver une géométrie similaire à celle du site de fixation des Nanofitines et identifier des charpentes humaines compatibles.
4. Transférer le site de fixation de la Nanofitine sur la charpente pour générer un nouveau ligand d'origine humaine doté d'une reconnaissance spécifique de la cible ainsi que de nouvelles propriétés intrinsèques.

Affinité et spécificité de 8 Nanofitines anti-GFP

→ Forte spécificité envers leur cible → Fortes affinités: $K_D < 10$ nM

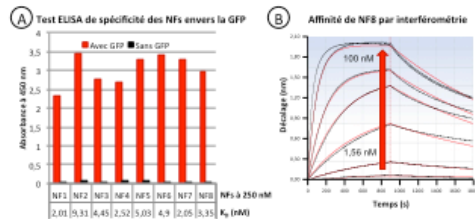


Figure 3: Mesures de spécificité et d'affinité des Nanofitines anti-GFP NF1 à NF8. A) Test ELISA des Nanofitines avec Hi-tag en présence (rouge) ou absence (noir) de GFP en fond de puits. B) Exemple de détermination d'affinité par interférométrie (Octet RED96, FortiBio).

Epitope binning → 2 épitoques distincts → NF1 ≠ NF2-8

Les sites de NF1 et NF7 ont été choisis pour le transfert sur de nouvelles charpentes.

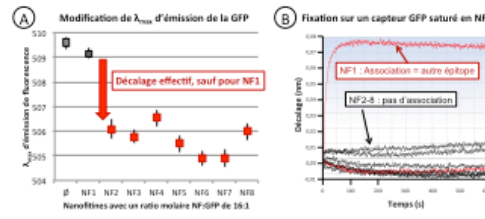


Figure 4: Distinction de deux sites de reconnaissance par les Nanofitines sur la GFP. A) Mesure de la longueur d'onde d'émission maximale de la GFP en présence de Nanofitine. B) Exemple de mesure de compétition par interférométrie.

Vers des Nanofitines humanisées: Greffe de site de fixation et production de charpentes protéiques humaines

Grâce au criblage de la PDB, de nombreuses charpentes humaines apparaissent comme **potentiellement compatibles** avec la greffe d'un site de reconnaissance issu d'une Nanofitine générée *in vitro* (géométrie du squelette carboné, orientation des chaînes latérales, accessibilité...). Il devrait donc être possible **d'accéder à des propriétés nouvelles**, tout en **conservant l'activité de fixation de la Nanofitine d'origine**. Cependant, ces nouvelles charpentes doivent répondre à un cahier des charges précis, notamment en termes d'expression en hôte bactérien ou de stabilité. Les travaux en cours portent sur la **production sous forme soluble de 3 charpentes d'origine humaine** ayant reçu les sites de fixation de NF1 ou NF7. Afin d'évaluer la **réussite du transfert**, leur caractérisation sera effectuée parallèlement à celle d'un homologue de Sac7d et sera comparée à celle des Nanofitines (Fig. 3 et 4).



Aperçu du poster: Huet S., Cinier M., Sanejouand Y-H. (2014). **Transfert de reconnaissance moléculaire : de la Nanofitine vers l'humain**. Journées Scientifiques de l'Ecole Doctorale Biologie-Santé, 9-10 Décembre 2014.

RECONNAISSANCE MOLÉCULAIRE: DE L'IN SILICO À L'IN VITRO ET VICE VERSA

En entreprise innovante



Nom: Affilogic SAS
Savoir-faire: Génération de Nanofitines®
Encadrant de thèse: Mathieu CINIER

Je suis **Simon HUET**, doctorant en 2^{ème} année de thèse CIFRE.



Je pratique la Biochimie et la Biologie moléculaire ...

... à l'aide de la Bioinformatique !

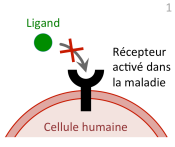
En laboratoire académique



Nom: UFIP, UMR CNRS 6286
Savoir-faire: Modélisation et Prédiction
Directeur de thèse: Yves-Henri SANEJOUAND
Ecole doctorale: Biologie Santé (ED 502)

Nous voulons développer de nouveaux médicaments, les Nanofitines®

1



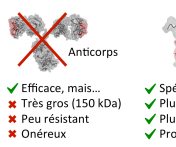
Ligand

Récepteur activé dans la maladie

Cellule humaine

Pour bloquer une interaction récepteur-ligand ...

2



Anticorps

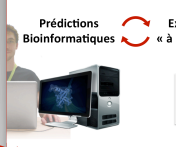
Nanofitine®

- ✓ Efficace, mais...
- ✗ Très gros (150 kDa)
- ✗ Peu résistant
- ✗ Onéreux

- ✓ Spécificité contrôlée *in vitro*
- ✓ Plus petit (10 kDa)
- ✓ Plus stable (pH 0-13, T_m=80°C)
- ✓ Production peu coûteuse (E. coli)

... À l'aide de protéines d'affinité alternatives aux anticorps, les **Nanofitines®** ...

3




Prédictions Bioinformatiques

Expériences « à la pailasse »

... Nous générons des ligands affins et spécifiques comme inhibiteurs ...

4



... Afin de créer des médicaments innovants et abordables !

De nombreuses pathologies impliquent des interactions protéine-protéine entre un récepteur à la surface des cellules et un ligand extracellulaire. Les signaux propagés par ces voies de signalisation peuvent être imités par des protéines affines et aux propriétés de neutralisation qui ainsi traitent la maladie.

Les protéines neutralisantes sont majoritairement des anticorps ou dérivés. Malgré leur efficacité, les anticorps souffrent de limitations notamment liées à leur structure complexe avec modifications post-traductionnelles (hétérogénéité, production onéreuse, faible stabilité) et à leur mode d'administration principalement systémique. L'alternative développée par Affilogic consiste à utiliser les Nanofitines®, des protéines à la charpente plus simple et plus stable (donc plus faciles d'utilisation, avec des voies d'administration alternatives) qui possèdent des affinités et spécificités semblables à celles des anticorps.

Le but de ce projet est d'optimiser la génération des ligands d'Affilogic: prédire leurs propriétés, les diriger contre des cibles difficiles à utiliser *in vitro* (par ex: des récepteurs cellulaires), etc. La synergie entre les prédictions *in silico* et leur vérification *in vitro* est également l'occasion de progresser vers une meilleure modélisation des propriétés physicochimiques des protéines et de leurs interactions sur le plan théorique.

Nous bénéficions d'une technologie attractive développée à la plateforme Affilogic (rapidité du procédé, exclusivité des produits) et d'un savoir-faire de modélisation sans cesse perfectionné à l'UFIP. Grâce à ces outils, nous désirons amener sur le marché de nouveaux médicaments aux modes d'administration les moins invasifs possible, à faible coût de production.


Ce que nous attendons de la modélisation et sa validation *in vitro* :

5



Cibler finement des protéines d'intérêt thérapeutique et mieux modéliser les interactions en jeu

6



Développer savoir et savoir-faire reconnus: publications et congrès en biotechnologies ...

De nombreux progrès sont réalisés dans le domaine de la bioinformatique. Nous désirons continuer à alimenter ces connaissances et profiter de ces avancées pour perfectionner les Nanofitines®. Notre objectif majeur est d'optimiser leur capacité d'inhibition et de cibler avec précision leur site d'interaction pour générer des agents de neutralisation de plus en plus performants.

Les retombées scientifiques de ces travaux seront d'une part l'amélioration des connaissances théoriques et pratiques dans le domaine des interactions protéine-protéine, et d'autre part leur communication via des articles scientifiques et des participations à des congrès en biologie et bioinformatique.

Le projet est rendu possible grâce à la participation de :

7



UNIVERSITÉ DE NANTES

8



Financements

Cette thèse est cofinancée par l'Union Européenne, l'Europe s'engage en Pays de la Loire avec le Fonds européen de développement régional.

Savoir-faire, équipements, réactifs, locaux à Nantes (France)

Les enjeux économiques, sociétaux et industriels

9



INFLAMMATION

OPHTALMOLOGIE

SYSTÈME NERVEUX CENTRAL

MALADIES INFECTIEUSES

Optimiser nos molécules pour aboutir à de nouveaux médicaments

10



Propriété intellectuelle

BREVET

Faire du licensing et breveter de nouvelles molécules/techniques

A terme, nous souhaitons que les Nanofitines® développées grâce à ces travaux servent à traiter des pathologies pour lesquelles les traitements actuels sont insuffisants (administration contraignante, etc.). Les enjeux de ce projet sont donc d'optimiser l'efficacité de nos molécules pour convaincre des partenaires de s'engager en clinique (Phase I) avec elles.

L'avancée des tests pré-cliniques et cliniques sera accompagnée par le dépôt de brevets pour l'entreprise et/ou le laboratoire académique, aboutissant à du licensing. Cette propriété intellectuelle pourra être protégée tant sur l'aspect génération inventive de molécules que sur les produits innovants en eux-mêmes.



affilogic
www.affilogic.fr



UFIP
ufip.univ-nantes.fr



www.simon-huet.fr



Man of Still training ...

- **Compétences au service de la thèse :**
Ma formation universitaire et une sensibilité aux outils informatiques m'ont amené à ce projet de thèse exploitant et perfectionnant cette double compétence biotechnologies et bioinformatique. Mon cœur de compétences s'articule donc autour des techniques d'ingénierie des molécules du vivant (du gène à la protéine), de la prédiction à la validation expérimentale.
- **Vers une carrière en industrie biotech/pharma :**
Ce travail de recherche en entreprise et en laboratoire académique m'a permis de préciser mon projet post-thèse: encadrer et faire de la recherche en industrie (R&D biotechnologies et pharma).

Aperçu du poster: Huet S. (2014). *Reconnaissance moléculaire: de l'in silico à l'in vitro et vice versa*. Les Doctoriales, 23-28 Mars 2014.

Thèse de Doctorat



Simon HUET

Reconnaissance moléculaire : de l'*in silico* à l'*in vitro* et vice versa

Molecular recognition : from *in silico* to *in vitro* and vice versa

Résumé

La reconnaissance moléculaire, ou formation spécifique de complexes entre molécules par l'intermédiaire d'interactions physico-chimiques, est un mécanisme d'action fondamental des protéines appliquées à un usage médical. Dans cette étude, nous explorons l'ingénierie rationnelle de cette fonction avec comme partenaire une nouvelle classe de protéines destinée à développer les médicaments de demain: les Nanofitines (NF). Dans l'objectif de guider les approches expérimentales (*in vitro*) par celles prédictives (*in silico*) dans une collaboration réciproque, ce manuscrit décrit successivement deux preuves de concept appliquées à des NF dirigées contre la protéine de fluorescence verte (GFP). La première approche, conçue pour l'humanisation de NF par transfert sur des protéines humaines, a consisté à isoler 11 résidus fonctionnels à greffer à partir de NF anti-GFP générées par *ribosome display*. Le squelette du feuillet beta exposant ces acides aminés clés a ensuite été recherché dans une banque structurale pour identifier trois protéines humaines hôtes arborant un feuillet similaire ($RMS \leq 0,63\text{\AA}$). Une fois greffées des résidus clés des NF, ces protéines ont démontré une insolubilité ou une absence de fixation détectable à la GFP, contrairement à un homologue structural des NF. La seconde approche a visé à réaliser le design *de novo* de NF anti-GFP par modélisation moléculaire. Cette démarche, proposée comme alternative rationnelle complémentaire du procédé de sélection *in vitro*, a abouti à la découverte d'une NF dont l'épitope ciblé s'avère compatible avec les prédictions réalisées. Cette stratégie nécessitera d'être complétée afin d'optimiser l'affinité des NF générées *de novo*.

Mots clés : Nanofitine, Ingénierie des protéines, Interaction protéine-protéine, Évolution dirigée, Modélisation moléculaire, Design rationnel *de novo*, Greffe de charpente, Humanisation protéique

Abstract

Molecular recognition, or specific complex formation between molecules through physico-chemical interactions, is one of the fundamental mechanisms of action of proteins intended for medical use. In this study, we have explored the rational engineering of this function involving a new class of proteins doomed to be part of the drugs of tomorrow: the Nanofitins. In order to guide experimental approaches (*in vitro*) by the predictive ones (*in silico*) in a mutual collaboration, this manuscript describes two successive proofs of concept applied to Nanofitins specifically directed towards green fluorescent protein (GFP). The first approach was designed to generate humanized Nanofitins following the transfert of their scaffold onto the surface of human proteins, and involved the isolation of 11 functional residues to be grafted from anti-GFP Nanofitins generated by ribosome display. The backbone of the beta-sheet displaying these key amino acids was then sought in a structural protein data bank to identify three human protein hosting a similar sheet ($RMS \leq 0.63\text{\AA}$). Upon grafting of Nanofitin's key residues, these proteins demonstrated insolubility or lack of detectable binding to GFP, unlike a structural homologue of Nanofitins. The second approach aimed to perform the *de novo* design of anti-GFP Nanofitins by molecular modeling. This method, proposed as a rational alternative complementing the current *in vitro* selection process, has led to the discovery of a Nanofitin whose targeted epitope appears to be consistent with computed predictions. This strategy will need to be completed in order to optimize the affinity of the *de novo* generated Nanofitins.

Key Words : Nanofitin, Protein engineering, Protein-protein interaction, Directed evolution, Molecular modelling, Rational *de novo* design, Protein scaffold grafting, Protein humanization