



École Doctorale INTERFACES
Approches interdisciplinaires: fondements, applications et
innovations

Titre de la thèse

THÈSE

présentée et soutenue publiquement le XXX

pour l'obtention du

Doctorat de l'Université Paris-Saclay

spécialité: Les sciences du vivant

par

M. David MIGNON

Composition du jury

Rapporteurs :	Dr. Prénom1 NOM1	Rapporteur externe
	Dr. Prénom2 NOM2	Rapporteur externe
	Pr. Prénom3 NOM3	Rapporteur interne
Examineurs :	Dr. Prénom4 NOM4	Examineur
	Dr. Prénom5 NOM5	Directeur de thèse
	Dr. Prénom6 NOM6	Directeur de thèse

Laboratoire de XXX

Remerciements

XXX

à XXX.

Table des matières

Liste des figures	ix
Liste des tables	xi
Abreviations	xiii
Introduction	1
1 CPD	5
2 proteus	7
3 Comparaisons d’algorithmes	9
3.1 Les méthodes pratiques	9
3.1.1 les protéines	9
Alignements Blast croisés	10
3.1.2 Description des tests	13
Ensemble «Tout actif»	13
L’ensemble «nombre d’actifs limité»	13
le choix des positions actives	14
positions en interactions	14
choix des positions actives	14
3.1.3 Définition de protocole comparable	15
Protocole heuristique	16
Protocoles Monte-Carlo	16
Seconde version de proteus	17
Protocoles "Replica Exchange"	17
Protocoles Toulbar2	19
3.1.4 Outils d’analyse des données	20

Table des matières

	Superfamily/SCOP	20
	Taux d'identité de séquences	20
	Taux d'identité par position	20
	Alignements Pfam	20
	Score BLOSUM	21
	similarité d'un ensemble à une famille Pfam	21
	Répartition de l'énergie selon les centiles	21
4	PDZ	23
4.1	Introduction	23
	23
	La fonction Rosetta	23
4.1.1	Description	24
4.2	Le modèle d'état déplié	24
4.2.1	Le maximum de vraisemblance des énergies de référence	24
	24
4.2.2	Recherche du maximum de vraisemblance	25
4.3	Méthodes de calcul	26
4.3.1	Fonction énergétique efficace pour l'état replié	26
4.3.2	Les énergies de référence de l'état déplié	27
4.4	Séquences expérimentales et modèles structuraux	28
4.5	L'ensemble des protéines PDZ	28
	Alignements Blast croisés	28
	similarité des homologues	28
	29
4.5.1	simulation Monte Carlo	30
4.5.2	Génération de séquence Rosetta	30
4.5.3	Caractérisation de la séquence	32
	32
4.6	Résultats	32
4.6.1	Structures et séquences expérimentales	32
	32
4.6.2	optimisation du modèle de l'état déplié	33
4.6.3	Évaluation de la qualité des séquences obtenues	33
	Tests de reconnaissance de famille	33
	Séquences et diversité de séquence	33

Scores de similarité Blosum	34
Tests de validation croisée	34
Conclusion	37
Bibliographie	45

Liste des figures

3.1	alignement de 1R6J et 2BYG obtenu avec Clustal Omega version 1.2.1 . . .	12
-----	--	----

Liste des tables

3.1	Les protéines	10
3.2	Pourcentage d'identité et e-value des alignements Blast native vs native pour nos protéine SH3.	10
3.3	Pourcentage d'identité et e-value des alignements Blast native vs native pour nos protéine PDZ (no= pas de touche avec une e-value inférieure à 10).	10
3.4	Pourcentage d'identité et e-value des alignements Blast native vs native pour nos protéine SH2.	11
3.5	Les protocoles heuristiques	16
3.6	Les protocoles Monte-Carlo	18
3.7	Les protocoles «Replica Exchange»	19
4.1	Les groupes d'acides aminés utilisés pour l'optimisation des énergies de référence.	27
4.2	La sélection de domaines protéiques PDZ	28
4.3	E-value et pourcentage d'identité des alignements Blast native versus native pour nos séquences PDZ.S'il n'y a pas de touche avec une E-value inférieure à 10,[] donne le pourcentage d'identité du couple dans l'alignement des 6 séquences sauvages.	29
4.4	Sélection des homologues.	29
4.5	Compositions en acides aminés des séquences expérimentales homologues aux positions enfouies et actives. pour les 8 protéines.	31
6	Les tests avec cinq positions actives	40
7	Les tests avec dix positions actives	41
8	Les tests avec vingt positions actives	42
9	Les tests avec trente positions actives	43

Abbreviations

H algorithme heuristique

MC algorithme Monte-Carlo

RE algorithme “Replica Exchange”

GMEC “Global minimal energie cost”

Pfam “Protein family databank”

Introduction

XXX

Contexte

XXX

XXX

Citation entre crochets [??].

Citation dans le texte ?.

Chapitre 1

CPD

Chapitre 2

proteus

Chapitre 3

Comparaisons d'algorithmes

3.1 Les méthodes pratiques

Nous cherchons maintenant à déterminer les performances et les qualités des différents algorithmes de proteus. Pour évaluer les différents algorithmes de proteus, comme pour leur établir un paramétrage, nous effectuons des séries de tests. Grâce à l'algorithme de type toulbar2 il est possible d'obtenir la séquence/conformation qui possède la plus haute énergie de dépliement. Cela constitue une information important qui va nous servir d'élément de comparaison. Le facteur temps est également un élément déterminant. Il est dans certain cas limitant, nous ne savons pas à l'avance quand toulbar2 termine. Et il apparaît d'emblée illusoire d'espérer voir ce programme converger dans toutes les situations intéressantes dans un temps raisonnable. D'autres métriques qui caractérisent les séquences d'acides aminés de meilleurs énergies obtenues seront également utilisées pour les évaluations et pour les paramétrages.

Dans la suite, on appelle «position active», une position pour laquelle, tous les types d'acides et tous les rotamères de chaque type d'acide aminé sont autorisés, au court de la recherche de proteus. On désigne «séquence/conformation» une séquence d'acides aminés munie à chaque position d'un rotamère (le backbone étant de toute façon fixé). Tandis ce que le terme simple «séquence» sans plus de précision désigne une séquence d'acides aminés.

3.1.1 les protéines

Les tests sont effectués sur neuf protéines choisies pour avoir des longueurs de backbone variées, plusieurs domaines représentés, mais aussi plusieurs structures pour chaque famille présente. Ainsi l'ensemble se décompose en deux protéines SH3 de 56 et 57 résidus, de trois protéines PDZ de longueur comprise entre 82 et 97 résidus et enfin de trois protéines

Chapitre 3. Comparaisons d'algorithmes

SH2 longues de 105 ou 109 résidus. L'ensemble a une moyenne, arrondie à l'unité inférieure, de quatre-vingt-neuf positions, voir les détails en table 3.1.

Code PDB	résidus	nombre de positions	domaine
1ABO	64-119	56	SH3
1CKA	134-190	57	SH3
1R6J	192-273	82	PDZ
1G9O	9-99	91	PDZ
2BYG	186-282	97	PDZ
1BM2	55-152	98	SH2
1O4C	1-105	105	SH2
1M61	4-112	109	SH2
1A81	9-117	109	SH2

Table 3.1 – Les protéines

Protein	1ABO	1CKA
1ABO	100 (6e-42)	26 (1e-07)
1CKA	26 (1e-07)	100 (2e-41)

Table 3.2 – Pourcentage d'identité et e-value des alignements Blast native vs native pour nos protéine SH3.

Protein	1R6J	1G9O	2BYG
1R6J	100(1e-59)	25 (3e-07)	no
1G9O	25 (3e-07)	100(2e-66)	35 (2e-11)
2BYG	no	35 (2e-11)	100(7e-71)

Table 3.3 – Pourcentage d'identité et e-value des alignements Blast native vs native pour nos protéine PDZ (no= pas de touche avec une e-value inférieure à 10).

Alignements Blast croisés

Protein	1BM2	1O4C	1M61	1A81
1BM2	100 (7e-74)	36 (2e-16)	38(6e-10)	35 (1e-13)
1O4C	36 (2e-16)	100(2e-79)	27(3e-10)	33 (2e-12)
1M61	38 (6e-10)	27 (3e-10)	100(6e-81)	57 (2e-47)
1A81	35 (1e-13)	33 (2e-12)	57(2e-47)	100(5e-83)

Table 3.4 – Pourcentage d'identité et e-value des alignements Blast native vs native pour nos protéine SH2.

align_clustalo_1R6Jvs2BYG.png

Figure 3.1 – alignement de 1R6J et 2BYG obtenu avec Clustal Omega version 1.2.1

3.1.2 Description des tests

Les tests sont répartis en deux ensembles :

1. un ensemble de tests où toutes les positions de la séquence sont actives (cela correspond aux situations de design complet de protéines)
2. un ensemble de tests où le nombre de positions actives est gardé sous contrôle de façon à maîtriser la taille de l'espace d'exploration

Ensemble «Tout actif» Pour le premier ensemble de tests, la totalité de la matrice d'énergie est exploitée et pour chaque position l'espace d'exploration correspond à l'espace d'état déclaré dans le fichier ".bb". C'est-à-dire que tous les types de résidu et tous les rotamères sont possibles à chaque position. Comme l'espace des séquences/conformations à explorer est gigantesque, nous ne faisons pas de tentatives de recherche du GMEC par méthode exacte.

Nous effectuons des recherches avec les algorithmes suivants :

- heuristique, noté H par la suite ;
- Monte-Carlo, noté MC ;
- «Replica Exchange», noté RE) ;

L'ensemble «nombre d'actifs limité» L'ensemble «Nombre d'actifs limité» est composé de six groupes de tests avec un nombre de positions actives fixe défini de la façon suivante :

1. aucune position active
2. une position active
3. cinq positions
4. dix positions
5. vingt positions
6. trente positions

Lorsqu'une position n'est pas active, l'acide aminé de la position est fixé en utilisant l'acide aminé de la séquence native. La chaîne latérale est, elle, laissée libre. Il n'y a donc jamais dans nos tests de position où l'état est complètement fixé.

Le groupe «aucune position active» n'est constitué que d'un test par algorithme pour chaque protéine. Il y a donc neuf tests par algorithme. Ce sont les tests pendant lesquels la séquence d'acides aminés est fixe et correspond à la séquence native de la protéine.

Pour les tests avec une seule position active, comme des temps de calcul le permettent, nous décidons d'être exhaustifs : Toutes les positions sont testées, il y a alors huit cent quatre tests par algorithme. Pour tous les autres groupes de tests (cinq,dix,vingt et trente positions actives), cinq tests sont effectués par protéine, c'est-à-dire quarante-cinq tests par algorithme.

le choix des positions actives Pour définir complètement les tests, il reste maintenant à décrire le choix des positions actives pour les groupes de numéro trois jusqu'à six. Il y a peu d'intérêt à tester des situations avec des positions actives sans interaction entre-elles. En effet, s'il existe une position active P dont chaque résidu est sans interaction avec tous les résidus possibles des autres positions actives, déterminer le meilleur état pour P est proche du test du groupe 2 avec P comme position active. Toutefois, cela n'est pas exactement la même question, parce que les positions actives différentes de P peuvent influencer la position de la chaîne latérale de positions inactives qui à leur tour peuvent influencer l'état de P. Ainsi, le choix des positions actives ne se fait non pas par tirage aléatoire, car le risque d'obtenir des positions avec peu d'interactions est trop grand. Il se fait sous contrainte d'interaction.

positions en interactions Pour cela, nous utilisons la notion de voisinage de proteus. Elle se définit de la façon suivante : Deux positions P et Q sont en interactions s'il existe un rotamère r_P de P et un rotamère r_Q de Q tels que :

$$|E(r_P, r_Q)| > S_{Vois}$$

avec S_{Vois} un seuil donné par l'utilisateur à la configuration de proteus (voir chap. ?? pour les détails).

Alors on appelle «n-uplet en interaction» la donnée de n positions avec $n \in \{5,10,20,30\}$ et d'un seuil S_{Vois} tels que pour toute paire de positions (P,Q) du n-uplet, P et Q sont en interactions.

choix des positions actives Pour définir les positions actives,nous exécutons proteus en mode verbeux, sans effectuer d'optimisation.Pour cela, il existe plusieurs façons de procéder, ici nous utilisons le mode Monte-Carlo avec une trajectoire de zéro pas. Ces exécutions produisent en sortie standard la liste des voisins pour chaque position au seuil donnée en paramètre. Pour chacune des neuf protéines, nous exécutons proteus avec S_{Vois} égal à dix, cinq et un à tour de rôle ; trois listes de voisins sont obtenues. Ensuite,un script dédié recherche dans ces listes, les n-uplets en interaction, en partant de la liste de voisins

au sens le plus fort, c'est-à-dire dix, vers celle au sens le plus faible (0.1). La recherche s'arrête lorsque cinq n-uplets au moins sont trouvés.

Nous obtenons quarante-cinq n-uplets pour le groupe à cinq (respectivement dix, vingt et trente) positions actives pour un seuil S_{Vois} égal à dix (respectivement dix, un et un). Les positions actives de tous les tests sont en annexe ??). Pour chaque n-uplet, un fichier de configuration de proteus est créé dans lequel la balise <Space_Constraints> fixe les positions inactives en utilisant le type d'acide aminé présent dans la séquence native.

3.1.3 Définition de protocole comparable

Nous voulons comparer les algorithmes très différents. Un algorithme peut garantir l'obtention du minimum global en énergie (GMEC) si l'exécution se termine, mais ne garantit pas qu'elle se termine. Un autre permet un contrôle très fin du temps d'exécution sans garantie du GMEC, et d'autres enfin ont des objectifs plus large que la seule obtention du GMEC. Mais le GMEC reste le meilleur point de commun. Nous allons donc y concentrer une part importante des comparaisons.

Nous devons noter également que l'obtention du GMEC est théorique, en pratique nous n'avons pas de preuve que le code de l'algorithme exact que nous utilisons n'a pas de bogue. Cependant, nous mettons de côté cette éventualité et dans toute la suite GMEC désigne aussi bien le minimum global en énergie que le résultat de toulbar2 lorsqu'il se termine. Le Monte-Carlo et le «Replica exchange» possèdent de nombreux paramètres de configuration, ce qui rend l'ensemble des protocoles possibles très grand. Se pose alors la question de l'optimisation du protocole. L'objectif fixé ici, n'est pas la recherche d'un protocole optimal pour chacun des tests, mais d'évaluer, avec les tests, un protocole optimisé par algorithme. Nous allons alors dans un premier temps, recherche les meilleurs paramétrages pour le Monte-Carlo et le «Replica Exchange» sur l'ensemble de tests «tout actif». Puis, sur la base des résultats obtenus, les protocoles seront fixés pour effectuer les comparaisons sur l'ensemble «tout actif» et celui à «nombre d'actifs limité». Le programme toulbar2 possède aussi de nombreuses options. Deux paramétrages différents seront utilisés.

Pour rendre les protocoles comparables, le temps d'exécution maximum est fixé à vingt-quatre heures pour tous les exécutions. Toulbar2 donne sa meilleure séquence/conformation en dernier, il n'y a donc pas post-traitement nécessaire. C'est également le cas pour le Monte-Carlo à condition de configurer l'impression de la trajectoire avec la balise *Print_Threshold* = 0. dans le fichier de configuration. Pour le "Replica Exchange" et l'heuristique, un tri des séquences selon l'énergie est nécessaire. Mais il n'y a pas beaucoup de séquences :

1. L' Heuristique fournit une séquence/conformation à chaque cycle.
2. Le "Replica Exchange avec *Print_Threshold* = 0 produit autant de fichiers de séquences/rotamères que de marcheurs. Chacun ne contenant pas plus de quelques dizaines de séquences/rotamères.

Nous pouvons donc négliger la durée du tri dans le temps total d'exécution.

Protocole heuristique Pour l'algorithme heuristique, il n'y a dans notre situation qu'un seul paramètre à renseigner : le nombre de cycles à effectuer. Quelques essais préliminaires sur la plus grosse protéine (Table 3.1) avec toute les positions actives, montre que la version utilisée de proteus peut effectuer jusqu'à environ 110000 cycles sur nos machines de calculs en l'espace de vingt-quatre heures. Ainsi, le protocole H est défini comme le protocole qui utilise le mode heuristique de proteus et qui effectue cent dix mille cycles. Sont également définis les variantes H-, H+ et H++ comme des protocoles plus courts ou plus longs à facteur entier près (Table 3.5). Par ailleurs, certaines comparaisons de l'heuristique avec le Monte-Carlo ont été faites avec une version précédente du programme proteus. Ce protocole sera noté h. Il diffère aussi de H par le fait que l'option d'optimisation du compilateur Intel utilisé est -O2 contre -O3 pour H.

Nom	nombre de cycles
H	110000
H-	1100
H+	330000
H++	990000
h	100000

Table 3.5 – Les protocoles heuristiques

Protocoles Monte-Carlo On distingue deux ensembles de protocoles Monte-Carlo. Dans le premier, les noms sont de la forme "mc*". Il rassemble les protocoles utilisés pour le paramétrage du Monte-Carlo. Le second est constitué des protocoles utilisés lors des comparaisons.

Les éléments à paramétrer pour l'algorithme Monte-Carlo sont les suivants :

1. la température
2. le nombre de pas (avec le nombre de trajectoires et la longueur de trajectoire)
3. Le seuil de voisinage
4. Les probabilités de changements de la séquence/conformation

Ce qui représente un ensemble de protocoles trop grand pour une approche exhaustive. Pour l'essentiel, nous allons faire varier les paramètres un par un, en prenant comme point de départ un protocole qui rend le comportement de marcheur Monte-Carlo «proche» de l'heuristique.

La température est le paramètre principal du Monte-Carlo, c'est elle qui contrôle le taux d'acceptation du critère de Metropolis. Alors, la première étape de cette optimisation va consister à faire varier la température, entre 0.001 et 0.5, en conservant les autres paramètres fixés (protocoles de mc0 à mc5). Le nombre de pas total effectué est le produit de deux paramètres, le nombre de trajectoires et la longueur de trajectoire. Les protocoles mc1b et mc2b testent l'effet d'une augmentation du nombre de pas. Tandis que mc2c et mc2d testent l'effet de la variation du nombre de trajectoires par rapport à la longueur. Le protocole mc2e s'intéresse aux probabilités de changement de la trajectoire. Il y a cinq balises dans proteus qui contrôle ces changements :

<Prot> donne la probabilité de modifications de rotamère à une position.

<Prot_Prot> donne la probabilité de modifications de rotamère à deux positions.

<Mut> donne la probabilité de modifications de type de résidu à une position.

<Mut_Prot> donne la probabilité de modifications de rotamères à deux positions.

<Mut_Mut> donne la probabilité de modifications de type de résidu à deux positions.

La table 3.6 donne les probabilités utilisées par ces cinq paramètres dans l'ordre de la liste précédente.

Enfin, mc4b se distingue des autres par un seuil de voisinage plus grand ((Table 3.6)).

Seconde version de proteus Pour la partie comparaison avec les autres algorithmes, quatre protocoles sont utilisés. Les protocoles MCa et MCb s'inspirent fortement de mc2d et mc2e, en étant adapté à la contrainte du temps de calcul de la comparaison et en utilisant la nouvelle version de proteus (les lettres capitales dans le nom des protocoles signifient l'utilisation de la dernière version de proteus). MCa- est une variante de MCa avec une trajectoire six fois plus courte. Enfin, MC0 s'inspire de mc0 dans le sens où la température est suffisamment froide pour que nous puissions considérer qu'il n'y a pas de baisse de l'énergie au cours d'une trajectoire.

Protocoles "Replica Exchange" L'algorithme «Replica Exchange» (RE) est une extension du Monte-Carlo. Les paramètres d'un protocole RE sont ceux d'un protocole Monte-Carlo plus trois autres :

Chapitre 3. Comparaisons d'algorithmes

Nom	Temp	Long. de trajectoire(mega)	Nb de trajectoires	Voisin	Proba
mc0	0.001	3	1000	10	0 ; 1 ; 0.1 ; 0;0
mc1	0.1	3	1000	10	0 ; 1 ; 0.1 ; 0;0
mc2	0.2	3	1000	10	0 ; 1 ; 0.1 ; 0;0
mc3	0.3	3	1000	10	0 ; 1 ; 0.1 ; 0;0
mc4	0.5	3	1000	10	0 ; 1 ; 0.1 ; 0;0
mc5	0.7	3	1000	10	0 ; 1 ; 0.1 ; 0;0
mc1b	0.1	6	1000	10	1 ; 1 ; 1 ; 1;0
mc2b	0.2	6	1000	10	0 ; 1 ; 0.1 ; 0;0
mc2c	0.2	3	10000	10	0 ; 1 ; 0.1 ; 0;0
mc2d	0.2	3000	1	10	0 ; 1 ; 0.1 ; 0;0
mc2e	0.2	3	1000	10	1 ; 0 ; 0.1 ; 0;0
mc4b	0.5	10	100	10	0 ; 1 ; 0 ; 1;0
MC0	0.01	1000	1	10	1 ; 0 ; 0.1 ; 0;0
MCA	0.2	6000	1	10	1 ; 0 ; 0.1 ; 0;0
MCA-	0.2	1000	1	10	1 ; 0 ; 0.1 ; 0;0
Mcb	0.2	6000	1	10	0 ; 1 ; 0.1 ; 0;0

Table 3.6 – Les protocoles Monte-Carlo

- le nombre de marcheurs
- la température pour chaque marcheur
- la période de «swap», c'est-à-dire la période (en nombre de pas) à laquelle le test de Hasting sur l'échange de température est effectué.

Pour avoir des exécutions en parallèle avec au plus un marcheur par coeur du processeur, nous limiter nos tests à quatre ou huit marcheurs. La distribution des températures est un élément déterminant dans le comportement des marcheurs, car c'est elle qui pilote en grande partie le taux d'acceptation des échanges de températures. Nous suivons l'idée proposée par Kofke de lui faire suivre une progression géométrique ($\frac{T_i}{T_{i+1}} = C$, avec C une constante) [???]. Ceci garantie alors que le taux d'acceptation d'échange entre T_i et T_{i+1} soit égale pour tout nos i. De plus, nous souhaitons centrer approximativement, nos distributions sur la température ambiante (environ 0.6 kcal/mol). Dans toute la suite, les températures et les énergies sont exprimées en kcal/mol.

Voici les températures pour le RE quatre marcheurs :

- 10, 1, 0.1 et 0.01
- 2, 1, 0.5 et 0.25
- 1, 0.5, 0.25 et 0.125

,et celles pour le RE huit marcheurs :

- 3 , 2 , 1.333 , 0.888 , 0.592 , 0.395 , 0.263 et 0.175

— 10 , 3.16 , 1 , 0.316 , 0.1 , 0.0316 , 0.01 et 0.00316

Ici les protocoles ne se font qu’avec une seule trajectoire par marcheur. Et la contrainte du temps de calcul se comprend comme vingt-quatre heures de calculs cumulées sur tous les marcheurs. Ainsi les longueurs de trajectoire sont définies pour le RE à quatre marcheurs comme le quart d’une trajectoire MC, pour le RE à huit marcheurs comme le huitième.

La table 3.7 donne les probabilités utilisées par les cinq balises qui contrôlent les modifications de la séquence/conformation à chaque pas, dans l’ordre de la liste de la section 3.1.3.

Nom	marcheurs	Temp	Traj (mega)	seuil voisin	Proba	swap period (me
RE4a	4	10<->0.01	1500	10	1 ; 0 ; 0.1 ; 0 ; 0	7.5
RE4b	4	1<->0.125	1500	10	1 ; 0 ; 0.1 ; 0 ; 0	7.5
RE4c	4	2<->0.25	1500	10	1 ; 0 ; 0.1 ; 0 ; 0	7.5
RE8a1	8	10<->0.00316	750	0	1 ; 0 ; 0.1 ; 0 ; 0	2.5
RE8a2	8	10<->0.00316	750	10	1 ; 0 ; 0.1 ; 0 ; 0	2.5
RE8b1	8	3<->0.175	750	10	0 ; 1 ; 0.1 ; 0 ; 0	7.5
RE8b2	8	3<->0.175	750	10	1 ; 0 ; 0.1 ; 0 ; 0	7.5
RE8b3	8	3<->0.175	750	10	1 ; 0 ; 0.1 ; 0 ; 0	1

Table 3.7 – Les protocoles «Replica Exchange»

Protocoles Toulbar2 Après avoir converti nos matrices au format «wcsp» grâce à un script dédié, nous pouvons utiliser toulbar2. Le protocole de recherche du GMEC est le suivant : L’exécutable toulbar2 de version 0.9.7.0 est lancé avec les options « -l=3 -m -d : -s », ce qui correspond au paramétrage conseillé dans la documentation CDP [??]. Si l’exécution se termine en moins de vingt-quatre heures, le protocole est achevé. Sinon le programme est arrêté et une seconde version (la 0.9.6.0) est lancée avec les options « -l=1 -dee=1 -m -d : -s ». Au bout de vingt-quatre heures si le programme n’est pas terminé, il est arrêté. La dernière séquence/conformation imprimée en sortie est collectée. Le choix de la seconde version et du paramétrage fait suite à une discussion avec monsieur Seydou Traoré.

Toulbar2 offre également la possibilité de fournir la liste des séquences/conformations dont l’énergie est comprise entre celle qui correspond au GMEC, E_{GMEC} et une autre E_{upper_bound} donnée en paramètre. Pour utiliser cette fonctionnalité nous utilisons le paramétrage : « -d : -a -s -ub= E_{upper_bound} ». Cependant, il s’avère que cette utilisation peut utiliser une quantité de mémoire vive importante. Alors, pour éviter tout plantage de nos machines, la mémoire que toulbar2 peut allouer est limité à 30 Go.

3.1.4 Outils d'analyse des données

Superfamily/SCOP Superfamily [?] est un ensemble composé :

- D'une base de données de modèles de Markov cachés, où chaque modèle représente une structure 3D d'un domaine de la classification SCOP.
- D'une série de scripts qui annotent à partir des informations de la base, les séquences données en entrée. Ici, nous utilisons uniquement l'association au modèle 3D le plus vraisemblable.

Nous travaillons avec la base de données à la version 1.75, et en conjonction, nous utilisons SAM (version 3.5) [?] et HMMER (version 3.0) [?] recommandés par l'équipe de Superfamily. Le paramétrage utilisé est celui par défaut.

Taux d'identité de séquences Soient S et N deux séquences d'acides aminés de même longueur l.

Le Taux d'identité $Id(S, N)$ de S par rapport N est égal au pourcentage de position où l'acide aminé est identique dans S et N. C'est-à-dire

$$Id(S, N) = \frac{\sum_{1 \leq i \leq l} \mathbb{1}(s_i, n_i)}{l} \times 100$$

avec s_i et n_i l'acide animé en i de S et de N respectivement, et $\mathbb{1}(x, y)$ la fonction qui vaut 1 lorsque $x=y$ et 0 sinon.

Taux d'identité par position Le taux d'identité d'un alignement A_S à la position i par rapport à une séquence N de même longueur se définit comme :

$$Id(A_S, i) = \frac{\sum_{1 \leq j \leq m} \mathbb{1}(s_i^j, n_i)}{m} \times 100, \text{ avec } m \text{ le nombre de séquences de } A_S.$$

Alignements Pfam Ce taux d'identité donne une mesure de la ressemblance entre un alignement et une séquence. Cela nous permet de comparer nos séquences calculées à la séquence native. Mais cela n'est pas notre seule objectif. Et nous voulons les évaluer par rapport à l'ensemble des séquences du domaine protéique de la native. La base de données Pfam (Protein families database) [?] regroupe les domaines protéiques connus en famille. Chaque famille étant représentée par des alignements multiples de séquences et des profils de modèles de Markov cachés [?]. Dans la suite, nous n'utiliserons l'alignement dit « seed » qui se base sur un petit ensemble de membres représentatifs de la famille et l'alignement « full », plus large, qui est généré par modèle de Markov caché à partir de l'alignement « seed ». Les alignements correspondent pour nous aux familles PF00017 (domaine SH2), PF00018 (domaine SH3) et PF00595 (domaine PDZ).

Score BLOSUM Pour tenir compte des ressemblances et des différences entre les acides aminés lors d'une substitution, nous avons besoin d'une matrice de coût. Nous utilisons la matrice BLOSUM62 (BLOCKS SUBstitution Matrix) [?] qui est construite à partir de blocs d'alignement très conservés (ici plus de 62% d'identités). Les fréquences des mutations y sont calculées. Le score BLOSUM d'une substitution est alors le logarithme de la fréquence de la mutation correspondante. À cela est ajouté un score de pénalités pour l'insertion d'un gap (c'est-à-dire un saut dans l'alignement).

On définit alors simplement un score de similarité de deux séquences de même longueur comme la somme des scores BLOSUM62 sur toutes les positions. De même le score de similarité d'un alignement par rapport à une séquence sera défini comme la moyenne des scores de similarité sur ensemble des séquences de l'alignement. Et enfin un score de similarité de deux ensembles de séquences alignés entre eux comme la moyenne des scores de similarité du premier ensemble par rapport aux séquences du second.

similarité d'un ensemble à une famille Pfam Afin de calculer un score de similarité d'un ensemble de nos séquences par rapport à une famille Pfam, il faut commencer par aligner nos séquences avec l'alignement de la famille. Pour cela nous utilisons le programme d'alignement BLAST [?]. Il implémente une heuristique qui recherche puis étend les meilleurs alignements locaux. Nous procédons comme suit :

1. La commande `blastpgp` est utilisée avec comme database (paramètre `-d`) l'alignement Pfam et comme séquence en entrée (paramètre `-i`) la séquence native.
2. Dans la sortie blast, la séquence qui produit l'alignement le plus significatif avec la native est collectée, notons-la S_0 .
3. L'alignement blast est alors utilisé pour positionner la native par rapport à S_0 et les gaps nécessaires pour aligner la native à S_0 sont ajoutés.
4. Le positionnement et les gaps sont alors appliqués tels quels à la liste de nos séquences.

Répartition de l'énergie selon les centiles Pour étudier différentes distributions d'ensemble de séquences/conformations selon l'énergie, nous déterminons les centiles de la façon suivante :

1. L'ensemble de séquences/conformations est trié selon l'énergie.
2. L'intervalle entre la meilleure énergie et la moins bonne est divisé en cent intervalles consécutifs contenant le même nombre de séquences/conformations (un centième du cardinal de l'ensemble).

3. Les quatre-vingt-dix-neuf valeurs d'énergie obtenues par ce découpage sont les centiles.

Chapitre 4

PDZ

4.1 Introduction

Nous cherchons maintenant à évaluer performances de notre modèle CPD sur un ensemble de protéines. Les domaines PDZ ("Postynaptic density-95 / Discs large / Zonula occludens-1") sont de petits domaines globulaires qui établissent des réseaux d'interactions entre protéines dans la cellule(1-6). Ils forment des interactions spécifiques avec des protéines cibles, généralement en reconnaissant quelques acides aminés à l'extrémité C-terminale. En raison de leur importance biologique, les domaines PDZ et leur interaction avec les protéines cibles ont été largement étudiées et utilisées en conception *in silico*. Des ligands ont été conçus pour moduler l'activité de domaines PDZ impliqués dans diverses pathologie (7-9).des domaines PDZ des ligands PDZ redessinés ont été utilisés pour élucider les principes du repliement des protéines et de l'évolution (10-13).Et ces domaines avec leurs ligands peptidiques fournissent des "benchmarks" pour tester les méthodes informatiques elles-mêmes (14-16).

A partir d'une sélection de domaines PDZ,nous optimisons les énergies de référence E/ref en utilisant un formalisme du maximum de vraisemblance. La performance du modèle est testée en générant des séquences par proteus pour chaque protéine de la sélection. Pour cela, nous utilisons les résultats précédents, en particulier ceux de la section (12), pour définir les valeurs des paramètres de l'optimisation du Monte-Carlo. Nous confrontons nos résultats à ceux de la fonction d'énergie de Rosetta , fonction, plus empirique que la notre, qui a connue le plus de succès(29-31).

La fonction Rosetta La fonction Rosetta (29-31) comprend un terme de répulsion de Lennard-Jones, un terme Coulomb, un terme de liaison hydrogène,un terme de solvation Lazaridis-Karplus (32) et des énergies de référence d'état dépliées. Il a un Un grand nombre de paramètres spécifiquement optimisés pour CPD, qui offrent des performances optimales,

mais une interprétation physique moins transparente que Proteus qui lui fournit la capacité de calculer les énergies libres formellement.

4.1.1 Description

La production de nos séquences calculées a été effectuée par des simulations Monte Carlo où toutes les positions de la chaîne polypeptidique ont été autorisées à muter librement, excepté celles occupées par une glycine ou une proline qui conservent leur type d'acide aminé. Ce qui nous procurent des milliers de variantes pour chaque domaine étudié. Nos tests comprennent une validation croisée où les énergies de référence optimisées sur un sous-ensemble de notre sélection sont utilisés sur un autre sous-ensemble de cette même sélection. Nous, réalisons également, une série de simulations Monte Carlo de deux domaines PDZ où le potentiel chimique hydrophobe des types d'acides aminés est progressivement augmenté, polarisant artificiellement la composition de la protéine. Comme le biais hydrophobe augmente, les acides aminés hydrophobes envahissent progressivement la protéine de l'intérieur, formant un noyau hydrophobe devenu plus grand que le naturel. La propension de chaque position du noyau à devenir hydrophobe à un niveau de biais plus ou moins élevé peut être considéré comme un indice d'hydrophobicité déterminé en fonction de la structure qui nous renseigne sur la propension du cœur à supporter des mutations.

4.2 Le modèle d'état déplié

4.2.1 Le maximum de vraisemblance des énergies de référence

L'énergie utilisée est ici est l'énergie de pliage de la protéine, c'est-à-dire la différence entre Ses énergies d'état pliées et dépliées (33). Un mouvement élémentaire possible est une "mutation", Nous modifions le type sidechain t à une position choisie i dans la protéine pliée, en assignant un rotamer r_0 particulier à la nouvelle chaîne latérale. Nous considérons la même mutation dans l'état déplié. Pour une séquence S particulière, l'énergie d'état dépliée est de la forme : $E^u = \sum_i E^r(t_i)$ La somme se fait sur tous les acides aminés de S , t_i représente le type à la position i .

Les grandeurs $E^r(t) \equiv E_t^r$ sont appelées "énergies de référence". Ils peuvent être considérés comme des potentiels. $-u = (E^f(...t'_i, r'_i...) - E^f(...t_i, r_i...)) - (E^r(t'_i) - E^r(t_i))$ avec f and u les changements d'énergie dans l'état plié et déplié.

Plus précisément, nous les choisirons telles qu'elles maximisent la probabilité des Séquences cibles. C'est à dire, nous celles qui sont le plus vraisemblable étant donnée l'observation des séquences cibles.

Soit S une séquence particulière. Sa probabilité de Boltzmann est :

$$p(S) = \frac{1}{Z \exp(-\beta_S)} \text{ où } S = G_S^f - E_S^u \text{ est l'énergie libre de repliement de } S, G_S^f \text{ est del'état replié, } \beta = \frac{1}{kT} \text{ est la température inverse et } Z \text{ une constante}$$

$$kT \ln p(S) = \sum_i E^r(t_i) - G_S^f - kT \ln Z = \sum_t n_S(t) E_t^r - G_S^f - kT \ln Z,$$

où la somme à droite se fait sur l'ensemble des types d'acides aminés et $n_S(t)$ est le nombre d'acide aminé de type t dans la séquence S .

Nous considérons maintenant un ensemble S de N séquences cibles \mathbf{S} ; On appelle \mathbf{L} la probabilité d'observer l'ensemble entier. \mathbf{L} est fonction des paramètres du modèle E_t^r . Comme nous voulons le maximum de \mathbf{L} sur les E_t^r , nous seré fère à \mathbf{L} comme leur vraisemblance.

Nous avons :

$$kT \ln \mathbf{L} = \sum_S \sum_i n_S(t) E^r(t) - \sum_S G_S^f - N kT \ln Z = \sum_t n_S(t) E_t^r - \sum_S G_S^f - N kT \ln Z,$$

avec $N(t)$ le nombre d'acide aminé de type t dans l'ensemble \mathbf{S} . Le facteur de normalisation Z (ou fonction de partition) est la somme sur l'ensemble les séquences possibles

\mathbf{R} :

$$Z = \sum_S \exp(-\beta_{\mathbf{R}}) = \sum_{\mathbf{R}} \exp(-\beta_{\mathbf{R}}^f) \prod_t \exp(\mathbf{R}(t) E_t^r) \text{ Pour maximiser } \mathbf{L}, \text{ nous considérons la dérivée de } Z$$

$$\frac{\partial}{\partial \mathbf{R}} \sum_{\mathbf{R}} n_{\mathbf{R}}(t) \exp(-\beta_{\mathbf{R}})$$

$$\mathbf{R} \exp(-\beta_{\mathbf{R}}) = \langle t \rangle.$$

La quantité à droite est la moyenne de Boltzmann du nombre $n(t)$ des acides aminés t sur toutes les séquences possibles. En pratique, c'est la population moyenne de t que nous voudrions obtenir dans une longue simulation Monte Carlo.

Pour que $\ln \mathbf{L}$ soit maximal il faut que ses dérivées par rapport à l' E_t^r soient nulles.

$$\frac{1}{N} \frac{\partial}{\partial E_t^r} \ln \mathbf{L} = \frac{1}{N} \sum_S n_S(t) - \langle t \rangle = \frac{N(t)}{N} - \langle t \rangle$$

$$\text{et donc } \mathbf{L} \text{ maximum} \Rightarrow \frac{N(t)}{N} = \langle t \rangle, \forall t \in aa$$

Ainsi, pour maximiser \mathbf{L} , nous devrions choisir E_t^r telle qu'une longue simulation donne les mêmes fréquences que la population moyenne.

4.2.2 Recherche du maximum de vraisemblance

Nous utilisons trois méthodes pour approcher les valeurs E_t^r .

1. La première consiste à avancer dans la direction du gradient de $\ln(\mathbf{L})$ en utilisant la règle itérative suivante (40) : $E_t^r(n+1) = E_t^r(n) + \alpha \frac{\partial}{\partial E_t^r} \ln(\mathbf{L}) = E_t^r(n) + (n_t^{exp} - \langle n(t) \rangle_n)$ avec α une constante, $n_t^{exp} = \frac{N(t)}{N}$ la population moyenne d'acide aminé de type t dans l'ensemble ciblé, $\langle \cdot \rangle_n$ la moyenne sur les séquences de l'ensemble.
2. La deuxième méthode est une variante de la première dans laquelle le n n'est pas constant, mais ajusté au cours de la simulation de la façon suivante. La règle (11) est utilisée trois fois avec trois valeurs différentes et constantes pour le ceci avec un jeu d'énergie de références identiques. une interpolation parabolique est effectuée sur les trois valeurs de la fonction proxy obtenues, le minimum de la parabole est calculée et est utilisée comme pour le cycle suivant, en terme duquel les énergies sont mises à jours.

3. La troisième méthode, utilisée précédemment (26,27), utilise une règle de mise à jour logarithmique :

avec kT l'énergie thermique, fixée empiriquement à 0,5 kcal/mol. Nous l'appelons la méthode logarithmique. Dans les dernières itérations, certaines valeurs ont tendance à converger lentement, avec des oscillations. Par conséquent, une règle modifiée où une énergie au cycle n et l'énergie au cycle $n-1$ sont moyennées avec un poids respectifs de 2/3 et 1/3.

Chaque itération, dans la suite, ont été effectuées avec 500 000 000 de pas par réplique de REMC.

4.3 Méthodes de calcul

4.3.1 Fonction énergétique efficace pour l'état replié

La matrice énergétique a été calculée avec la fonction d'énergie efficace suivante pour l'état plié : $E = E_{bonds} + E_{angles} + E_{dihe} + E_{impr} + E_{vdw} + E_{Coul} + E_{solv}$

Les six premiers termes de l'équation (13) représentent l'énergie interne de la protéine. Ils sont tirés de la fonction d'énergie empirique Amber ff99SB (42), légèrement modifiée pour le CPD.

Les charges du backbone ont été remplacées par un ensemble unifié, obtenu en faisant la moyenne sur l'ensemble des types d'acides aminés et ajuster légèrement pour rendre la partie backbone de chaque acide aminé neutre (43). Le dernier terme à droite de l'équation (13), E_{solv} , représente la contribution du solvant. Nous avons utilisé un modèle de solvant implicite "Generalized Born + Surface Area" ou GBSA (44) :

$$E_{solv} = DGB + E_{surf} = \frac{1}{2} \left(\frac{1}{\epsilon_W} - \frac{1}{\epsilon_P} \right) \sum_{ij} q_i q_j \left(r_{ij}^2 + b_i b_j \exp \left[-\frac{r_{ij}^2}{4b_i b_j} \right] \right)^{-\frac{1}{2}} + \sum_i \sigma_i A_i$$

Ici, ϵ_W et ϵ_P sont les constantes diélectriques du solvant et de la protéine; r_{ij} est la distance entre les atomes i et j .

σ_i est un paramètre qui représente la préférence de chaque atome à être exposé ou caché du solvant. Les atomes du soluté sont divisés en quatre groupes avec pour chacun un en cal/mol/Å² :

non polaire -5

aromatique -40

polaire -80

ionique -100

ON attribue aux atomes d'hydrogène un coefficient de surface de 0. Les surfaces sont calculées par l'algorithme de Lee et Richards (46), qui est implémenté dans le programme

XPLOR, en utilisant un rayon de “????” de 1,5 Å. Les simulations MC utilisent une constante diélectrique = 4 ou 8.

Dans le terme énergétique GB, le rayon de solvation atomique B_i approxime la distance de la surface à évaluer la surface en fuite, un facteur est appliqué aux zones de contact des chaînes latérales simplifiées. D

4.3.2 Les énergies de référence de l'état déplié

Dans le modèle CPD, l'énergie de l'état déplié dépend de la composition de la séquence par l'ensemble des énergies de référence E_t^r (équation 1). Ici, les énergies de référence ont été attribuées en f

Par conséquent, le modèle devrait être transférable dans une famille de protéines. Distinguer les positions enterrées / exposées double le nombre de paramètres E_t^r à ajuster. Inversement, pour réduire

$$E_t^r = E_c^r + t^r$$

avec E_c^r est un paramètre ajustable, tandis que t^r est une constante, calculée comme la différence d'énergie même avec le solvant.

Groupe	acides aminés	propriétés
1	Ala, Cys, Thr	petit
2	Ser	
3	Glu, Asp	chargé négativement
4	Gln, Asn	polaire
5	Ile, Leu, Val	apolaire
6	Met	non polaire
7	Hip, Hid, Hie	chargé positivement
8	Arg	
9	Lys	
10	Phe, Trp	aromatique
11	Tyr	
12	Gly, Pro	non mutable

Table 4.1 – Les groupes d'acides aminés utilisés pour l'optimisation des énergies de référence.

Plus précisément, nous effectuons des simulations MC d'un peptide étendu (le peptide Syndecan1 ; voir Ci-dessous) et calculons les énergies moyennes pour chaque type d'acide aminé à chaque position peptidique (à l'exclusion des positions terminales). Nous prenons les différences entre les types d'acides aminés et les moyennons sur les positions peptidiques.

Pendant, la maximisation de la vraisemblance, E_c est optimisé tandis que t est fixe. Pour optimiser les) avec des fréquences cibles correspondant aux fréquences expérimentales soit des classes d'acides ami

4.4 Séquences expérimentales et modèles structurels

4.5 L'ensemble des protéines PDZ

Nous sélectionnons huit protéines de la famille PDZ dont les structures cristallines sont connues, avec les trois présentes dans l'ensemble étudié au chapitre précédent : 1G9O, 1R6J et 2BYG, aux quelles sont ajoutés 1IHJ, 1N7E, 3K82 et Cask, Tiam1 représenté par et ... dans la base de données PDB. Cela constitue un ensemble où le nombre de positions actives, c'est à dire les positions qui vont être muté, est du même ordre pour chaque séquence d'acide aminé des protéines (voir le tableau 4.2).

Code PDB	résidus	nombre de positions actives
1G9O	9-99	76
1IHJ	13-105	82
1N7E	668-761	79
1R6J	193-273	72
2BYG	186-282	82
3K82	305-402	80
Cask	487-568	74
Tiam1	838-930	84

Table 4.2 – La sélection de domaines protéiques PDZ

Alignements Blast croisés Pour caractériser les homologies dans cet ensemble, une série de requête blast est effectuée sur chaque paire de séquences en utilisant le programme blastp avec les options comme indiqué en (45). Il apparaît que 1R6J et TiAM1 sont atypiques dans l'ensemble avec, aucun homologue avec une E-value inférieure à $1e-7$ et plusieurs E-value supérieur à 10. 3K82 est la protéine plus consensuelle, ayant d'une part une homologie avec toutes les autres à au plus $6e-04$, et d'autre part ayant 4 homologues à moins de $2e-10$, pour pourcentage d'identité compris entre 30 et 46. Globalement, il n'y a que peu d'homologies, la plus forte n'étant que de $3e-15$ entre 3K82 et 2BYG pour un pourcentage d'identité de 37. Les détails sont dans le tableau 4.3.

similarité des homologues Pour définir les fréquences d'acide aminés cibles pour maximiser nos vraisemblances, nous sélectionnons un ensemble de séquences homologues pour chacune de nos 8 protéines. Pour cela, nous effectuons des recherches blast avec comme requête la séquence extraites du fichier PDB sur la base de données "siwwprot + trEmBL" d'Uniprot avec la matrice BLOSUM62 sans l'option « filtre » et avec l'option

4.5. L'ensemble des protéines PDZ

Protein	1G9O	1IHJ	1N7E	1R6J	2BYG	3K82	CASK	TIAM1
1G9O	2e-66 (100)	5e-10 (40)	0.002 (25)	3e-07 (25)	2e-11 (35)	1e-12 (30)	5e-05 (25)	9e-07(35)
1IHJ	5e-10 (40)	3e-68 (100)	2e-07 (27)	[18]	2e-08 (27)	9e-14 (46)	4e-06 (35)	[16]
1N7E	0.002 (25)	2e-07 (27)	3e-67 (100)	[21]	3e-14 (36)	2e-10 (37)	9e-12 (30)	5e-05 (35)
1R6J	3e-07 (25)	[18]	[21]	1e-59 (100)	[17]	1e-06 (32)	0.007 (32)	[18]
2BYG	2e-11 (35)	2e-08 (27)	3e-14 (37)	[17]	7e-71 (100)	3e-15 (37)	2e-07 (28)	5e-05 (41)
3K82	1e-12 (30)	9e-14 (46)	2e-10 (36)	1e-06 (32)	3e-15 (37)	4e-70 (100)	1e-07 (27)	6e-04(33)
Cask	5e-05 (25)	4e-06 (35)	9e-12 (30)	0.007 (32)	2e-07 (28)	1e-07 (27)	7e-61 (100)	5e-04(33)
Tiam1	9e-07 (35)	[16]	5e-05 (35)	[18]	5e-05 (41)	6e-04 (33)	5e-04 (33)	1e-68 (100)

Table 4.3 – E-value et pourcentage d'identité des alignements Blast native versus native pour nos séquences PDZ. S'il n'y a pas de touche avec une E-value inférieure à 10,[] donne le pourcentage d'identité du couple dans l'alignement des 6 séquences sauvages.

« Gapped ». Nous obtenons un premier ensemble pour chaque cas en se limitant aux homologues de bonnes qualité au regard de E-value et du pourcentage d'identité, tout en conservant en même temps une certaine diversité. Cela oblige pour certaines protéine à accepter des E-values plus haute que 1e-40, notamment 1IHJ et 1G9O , respectivement 1e-32 et 1e-10 ,pour avoir un nombre d'homologue suffisant. Ensuite, les redondances les plus flagrantes sont enlevées manuellement. Finalement, les ensembles se composent de 42 à 126 homologues, avec des pourcentages d'identité supérieurs à 66 % excepté pour 1IHJ où il a fallut descendre jusqu'à 38% d'identité. Voir le tableau pour les détails 4.4.

protéines	% identité		
1G9O	62	1e-32	67-95
1IHJ	42	1e-10	38-95
1N7E	48	1e-45	84-95
1R6J	85	1e-43	85-95
2BYG	43	1e-41	78-95
3K82	50	1e-46	81-95
Cask	126	7e-28	60-85
Tiam1	50	2e-23	60-85

Table 4.4 – Sélection des homologues.

Pour chaque ensemble d'homologues, notons le H , nous calculons la moyenne sur toutes les séquences et toutes les positions pour obtenir des fréquences globales d'acides aminés. Les fréquences sont déterminées séparément pour les positions enfouies et exposées. Notons les $f_t^b(H), f_t^e(H)$, où t représente un type d'acide aminé et b et e représentent respectivement les mêmes moyennés, ce qui donne deux ensembles distincts de fréquences d'acides aminés f_t^b et f_t^e pour

Pour réaliser les calculs Monte Carlo, les structures ont été préparées et les matrices d'énergie calculées à l'aide d'une procédure décrites précédemment (15,50). Deux segments

manquants dans le domaine Tiam1 (résidus 851-854 et 868-869) ont été construits en utilisant le programme Modeller (51). Le ligand peptidique a été retiré de la structure PDB avant de calculer la matrice d'énergie. Pour chaque paire d'acides aminés, l'énergie d'interaction a été obtenue après 15 pas de minimisation de l'énergie, avec le backbone fixé et seulement les interactions de la paire entre les autres chaînes et le backbone. Cette courte minimisation simplifie l'approximation discret. Les rotamères de chaînes latérales utilisés sont une version légèrement étendue de la librairie de Tuffery et al (52), qui possède un total de 254 rotamères (sur l'ensemble des types d'acides aminés). Cette extension comprends des orientations d'hydrogène supplémentaires pour les groupes OH et SH (48). Cette bibliothèque de rotamères a été choisie pour sa simplicité et parce qu'elle a donné de très bonnes performances dans les tests de placement de chaînes en comparaison au programme spécialisé scwlr4 qui utilise une bibliothèque beaucoup plus grande (53,54).

4.5.1 simulation Monte Carlo

La conception des séquences est réalisée avec Proteus,

D'abord, pour optimiser les énergies de référence, nous faisons des simulations où environ la moitié des positions peuvent muter à la fois. Ensuite, les modèles optimisés ont été testés avec des simulations où tous les positions sauf celles natives en Gly ou Pro, sont libre de muter. C'est également le cas pour les simulations de la titration hydrophobe.

Dans les deux cas des mutations se sont produites au hasard, soumises uniquement à la fonction d'énergie MMGBSA qui entraîne la simulation. Les simulations Monte Carlo utilisent des mouvements à une ou deux positions, où les rotamères, les types d'acides aminés ou les deux peuvent changer. Pour les mouvements à deux positions, la deuxième position est choisie parmi celles qui avaient une énergie d'interaction significative avec la première (c'est à dire 10 kcal/mol ou plus). De plus, l'échantillonnage est amélioré par l'échange de réplique (REMC), où plusieurs simulations MC sont exécutées en parallèle, à différentes températures. Des échanges périodiques

4.5.2 Génération de séquence Rosetta

Des simulations Monte Carlo ont également été réalisées à l'aide du programme et de la fonction d'énergie Rosetta (37). Les simulations sont faites en utilisant la version 2015.38.58158 de la suite (librement disponible en ligne), en utilisant la commande :

```
Fixbb -s Tiam1.pdb -resfile Tiam1.res -nstruct 10000 -ex1 -ex2 -linmem_ig 10
```

où les options ex1 et ex2 activent une recherche améliorée des rotamères pour les chaînes latérales enfouies. La dernière option correspond au calcul de l'énergie à la volée,

4.5. L'ensemble des protéines PDZ

aa	1G9O	1IHJ	1N7E	1R6J	2BYG	3K82	cask	tiam1
ALA	5.8	10.4	14.5	8.3	12.1	4.6	4.6	7.1
CYS	3.0	1.6	0.1	2.6	0.2	0.4	3.0	0.0
THR	3.0	1.4	6.1	8.9	6.7	2.5	4.4	3.0
SER	4.2	8.5	3.2	7.2	1.8	7.1	4.4	4.8
GLU	7.4	1.4	0.0	0.3	0.1	6.3	6.3	5.9
ASP	6.3	3.1	5.9	0.2	8.0	2.4	3.9	3.0
ASN	2.9	0.3	2.9	3.3	3.9	2.4	0.7	2.9
GLN	3.2	3.0	0.0	0.7	1.1	4.7	1.4	0.1
ILE	7.0	22.1	23.4	17.0	13.3	3.3	19.7	11.6
VAL	25.8	16.4	7.9	18.8	18.6	1.8	13.8	13.1
LEU	17.2	13.6	29.9	14.6	18.8	5.3	15.1	25.5
MET	1.2	0.8	0.1	2.6	0.0	0.6	8.4	1.5
HID	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HIE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HIP	0.0	0.7	0.0	3.4	2.6	0.3	1.2	0.1
ARG	2.8	6.5	2.9	0.3	0.2	4.4	0.6	2.9
LYS	0.1	1.8	0.2	5.7	4.4	2.7	7.1	5.8
TRP	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0
PHE	6.4	6.8	0.2	4.2	3.08	4.3	3.9	5.9
TYR	3.4	0.3	2.8	1.1	2.6	5.4	0.0	5.7
GLY	0.1	0.9	0.0	0.5	2.3	1.0	0.0	0.0
PRO	0.1	0.3	0.0	0.0	0.1	0.2	0.6	0.0

Table 4.5 – Compositions en acides aminés des séquences expérimentales homologues aux positions enfouies et actives. pour les 8 protéines.

et les paramètres par défaut sont utilisées pour les autres options. Les résidus Gly et Pro présents dans la protéine sauvage ne sont pas autorisé à muter, et les positions qui mutent ne peuvent pas le faire en Gly ou Pro (comme dans les simulations Proteus). Des simulations sont exécutées pour chaque domaine PDZ jusqu'à obtenir 10 000 séquences uniques de faible énergie, ce qui correspond à des durées d'exécution d'environ 5 minutes par séquence sur un seul coeur d'un processeur Intel récent, pour un total de 10 heures par protéines en utilisant 80 coeurs. C'est tout à fait comparable au coût des calculs Proteus , en comptant le temps de caluls de la matrice d'énergie plus celui des simulations Monte Carlo.

4.5.3 Caractérisation de la séquence

Les séquences calculées sont comparées à l'alignement Pfam pour la famille PDZ, en utilisant la matrice Blosun40 et une pénalité d'écart de -6; Cette matrice est appropriée pour comparer des homologues éloignés (séquences CPD et Pfam ici). Chaque séquence Pfam est également comparée à l'alignement Pfam, ce qui permet de comparer des séquences calculées et un couple de domaine PDZ naturels. Pour ces comparaisons Pfam/Pfam, si un domaine de test T fait partie de l'alignement, la comparaison T/T n'est pas prise en compte, pour être plus cohérent avec les comparaisons calculées/Pfam. L'alignement Pfam utilisée est "RP55", composée de 12255 séquences. Les similitudes sont calculées pour d'abord les 14 résidus du coeur et pour 16 résidus de surface définis par leur presque total enfouissement ou exposition (voir ??) et enfin pour l'ensemble des positions de la protéine.

Les séquences calculées sont soumises à la bibliothèque de modèle de Markov Caché Superfamily (57,58) qui tente de classer les séquences selon la base de données structurales de protéines SCOP (59). La classification étant basée sur la version 1.75 de SCOP et 3.5 des Superfamily (voir plus haut). Le programme hmmscan est exécuté avec un seuil de valeur E de 10^{-10} et une totale de 15438 modèles dans la base SCOP.

Pour comparer la diversité des séquences produites avec la diversité des séquences naturelles, nous utilisons l'entropie par position, à partir de la formule $S_i = -\sum_{j=1}^6 f_j(i) \ln f_j(i)$

avec $f_j(i)$ la fréquence du type de résidu j à la position i , où i est distingué les 20 types d'acides aminés, nous utilisons les résidus LVIMC, FYW, G, ASTP, EDNQ et KRH. Cette classification a été obtenue par une analyse de clusters sur les

4.6 Résultats

4.6.1 Structures et séquences expérimentales

Les structures tridimensionnelles (3D) des quatre domaines PDZ de test sont représentées sur la Fig. 1A. Quatorze résidus de noyau (identifiés visuellement) des différentes structures se superposent bien. Tandis que les boucles et les chaînes terminales affichent de grands écarts. L'hélice $\alpha 2$ de Tiam1 est tournée légèrement vers l'extérieur par rapport aux trois autres structures (70). Les écarts sont dus à la divergence des identités des paires de séquences. Les écarts se situent entre 1,0 et 2,1 de l'identité des séquences.

La conservation des séquences dans les quatre domaines PDZ et un sous-ensemble de l'alignement "seed" de Pfam est représenté sur la figure 2. Les 14 positions utilisées pour définir le noyau hydrophobe sont bien conservées dans l'alignement des "seed" de Pfam, mais pas totalement. L'Arg, Lys et Gln apparaissent à certaines positions, puisque dans de

petites protéines comme des domaines PDZ, La longue partie hydrophobe de ces chaînes latérales peut être enfui dans le noyau tout en permettant à la pointe polaire de la chaîne d'être exposé au solvant. Quelques résidus Asp et Glu apparaissent aussi, dans les endroits où l'alignement des séquences peut ne pas très bien refléter la superpositions 3D les chaînes latérales.

4.6.2 optimisation du modèle de l'état déplié

Nous optimisons les énergies de référence E_i pour les six protéines, en utilisant leur homologues naturels. La classe dépend par construction du i défini dans pour chaque classe, qui ont été calculés avec la mécanique quantique pour huit types possibles (les positions Pro et Gly étant fixées).

4.6.3 Évaluation de la qualité des séquences obtenues

Tests de reconnaissance de famille Les simulations Proteus utilisent l'algorithme Monte Carlo avec échange de réplique (REMC) avec huit répliques et 750 millions de pas par réplique, avec des énergies thermiques kT qui varient de 0,125 à 3 kcal/mol. Toutes les positions sont autorisés à muter librement dans tout les types d'acides aminés excepté Gly et Pro. Les simulations ont été faites avec la fonction d'énergie MMGBSA, sans aucune introduction de biais vers les séquences naturelles ni aucune limite sur le nombre de mutations. Les 10 000 séquences avec les énergies les plus faibles parmi celles échantillonnées par au moins une des répliques MC sont retenue pour l'analyse. De la même façon, 10 000 séquences produites par Rosetta ont été retenue. Ces séquences sont analysées par les outils de reconnaissance de repliment "Superfamily" (58,71), voir tableau 4. Avec une constante diélectrique de 8, nous avons obtenu un pourcentage élevé de séquences correctement associées à la famille et superfamille PDZ : 91% pour Tiam1 et 100 pour Cask, avec des "E-values" d'environ 10^{-3} pour les affectations à la familles. Ces valeurs sont semblables à celles obtenues par Rosetta (90 et 98%).

Séquences et diversité de séquence Les séquences Tiam1 et Cask calculées par Proteus, par Rosetta et les séquences naturelles sont montrées à la figure 3. Pour les quatorze résidus du noyau et la figure 4 pour les seize résidus de surface (Tiam1 uniquement). Les séquences sont représentées par des logos de séquences. Comme on l'a vu dans de nombreuses études de CPD antérieures (30,72) l'accord avec l'expérience pour les positions du coeur est très bon, alors que l'accord en ce qui concerne les résidus de surface est nettement plus faible. La diversité des séquences naturelles et des séquences calculées est caractérisé par la moyenne sur la séquence de l'exponentielle de l'entropie résiduel (voir Méthodes), ce qui correspond à un nombre moyen de classe de séquence échantillonnées

par position. Par exemple, une valeur de 2 à une position particulière indique que les acides aminés de deux des six classes sont présents à cette position au sein de l'ensemble des séquences analysées. Une valeur moyenne globale de deux indique qu'en moyenne, deux classes d'acides aminés sont présentes à n'importe quelle position dans les séquences analysées. Comme référence l'ensemble Pfam RP55 et 12 255 séquences naturelles a une entropie moyenne de 3,4. Le regroupement des séquences Tiam1 et Cask calculés donne une entropie de 2,2 avec Rosetta et 2,0 avec Proteus. Ce qui indique que ces deux seules géométries de backbone ne peuvent pas être attendre les mêmes niveaux de diversités que le grand ensemble RP55. Prenant les 10000 séquences Monte Carlo de meilleure énergie échantillonné à température ambiante (au lieu des 10 000 de plus basses énergies échantillonnées collectivement par toutes les répliques à toutes les températures) et la mise en commun de Tiam1 et Cask donne une entropie globale plus élevée de 2,9 avec Proteus. Pour Rosetta, l'entropie dans le noyau est seulement légèrement inférieur à la moyenne sur toutes les positions. Pour Proteus, c'est nettement inférieur (1,25). Pour les séquences Pfam-Rp55, cette entropie est de 1,8.

Scores de similarité Blosum La figure 5 montre les scores de similarité Blosum40 entre les séquences calculées et les séquences naturelles. Avec Proteus, pour Tiam1 et Cask, les similitudes globales se chevauchent au pied du sommet des scores naturels, sont comparables aux valeurs des séquences Rosetta. Pour les résidus de surface, montrés séparément, la similitude avec les séquences naturelles sont faibles (scores inférieur à zéro), à la fois pour Proteus et Rosetta. Avec $\epsilon_p = 8$, Proteus donne presque la même similitude moyenne sur toutes les positions Tiam1 et Cask.

Tests de validation croisée Comme premier test de validation croisée, nous utilisons les énergies de référence optimisées à l'aide des homologues de Tiam1 et Cask à deux entre domaines PDZ : DLG2 et syntenine. Les scores Superfamily sont comparables à ceux obtenus pour Tiam1 et Cask, avec 100% de reconnaissance de la famille (tableau 4). Les séquences calculées avec Rosetta pour la DLG2 et la syntenine ont également obtenues une reconnaissance de la famille dans 100% des cas. Comme validation croisée supplémentaire, nous optimisons les énergies de références en utilisant un ensemble alternatif de domaines PDZ : DLG2, syntenin, PSD95, GRIP, INAD et NHERF. Pour distinguer les variantes du modèle, nous nous référons à cette nouvelle variante en tant que modèle n=6 (il utilise six domaines PDZ pour la paramétrisation) et le modèle initial comme modèle n=2. Les énergies de référence nouveau n=6, sont alors utilisées pour produire des séquences Tiam1 et Cask, qui sont alors soumises aux tests Superfamily et des calculs de similarité. La performance de Tiam1 sur la super-famille est légèrement dégradée par

6 et WT sont superposés à la figure 6. Ce sont très semblables les uns aux autres, avec une déviation de 1,2 Å. Au cours de la trajectoire MD de seq-6, les écarts par rapport à sa structure MD moyenne fluctuent.

4.7 Application : Croissance du noyau hydrophobe

Comme application de nos modèles optimisés, nous examinons la possibilité de conception du coeur hydrophobe des domaines PDZ. Chaque domaine PDZ est soumis à une simulation REMC avec une succession de fonctions énergétiques biaisées qui favorisent de plus en plus les résidus hydrophobes. La première simulation comprend un terme d'énergie de biais $\delta = 0,4 \text{ kcal/mol (par position)}$ qui pénalise les types d'acides aminés hydrophobes (I, L, M, V, A, W, F et Y). La dernière $-0,4 \text{ kcal/mol (par position)}$ qui favorise les types hydrophobes. Les valeurs d'énergie de biais intermédiaire $\delta = 0,2 \text{ kcal/mol}$ et $\delta = -0,2 \text{ kcal/mol}$ sont également simulées. En diminuant progressivement la valeur du biais d'énergie.

Les résultats pour Tiam1 sont présentés à la figure 7. À la plus grande valeur δ le coeur hydrophobe de Tiam1 est -12 changements, reflétant le biais. Environ 2/3 des changements se produisent dans des éléments de structure.

Nous calculons également un paramètre pour décrire le nombre de changement relatif de type d'acide aminé par unité de l'énergie du biais. Ce paramètre est défini comme le nombre de positions de résidu changées de non-polaire à polaire, divisé par le produit de la variation dans l'énergie de polarisation et le nombre moyen N de position non polaires à biais nul. Nous l'appelons la sensibilité hydrophobe ψ_h . Pour le domaine PDZ Tiam1, ce calcul donne : $\psi_h = \frac{1}{N} \text{frac} = 0,88 \text{ changements par position et par kcal/mol}$. Pour Cask, la sensibilité hydrophobe est $\psi_h = 0,71 \text{ changements par position et par kcal/mol}$.

4.8 Discussion

4.8.1 limites du modèle

Nous paramétrons un modèle CPD simple pour la conception de domaine PDZ, adapté aux applications à conception de nombreuses séquences et mis en oeuvre dans le logiciel Proteus. Pour la modélisation de l'état replié, nous utilisons un champ de force protéique de haute qualité. Nous avons testé deux ensembles de paramètres d'énergie de surface atomique σ_i et deux modèles de solvant "GB généralisé", le modèle NEA et le exact GB, avec une constante diélectrique $\epsilon_P = 8$ pour le NEA, et $\epsilon_P = 4$ pour le exact GB quantités qui caractérisent le modèle de solvant et sont les principaux paramètres.

La représentation de l'état déplié utilise un modèle simple caractérisé par une ensemble de potentiels chimiques d'acide aminé empiriques ou énergies de référence. Ces énergies sont choisi par une procédure de maximisation de la probabilité, formulée ici, afin de reproduire la composition d'acides aminés d'homologues naturels soigneusement sélectionnés. L'état déplié utilisé ici est plus raffinée qu'auparavant (76), puisque des valeurs d'énergies de référence distinctes sont utilisés pour les positions d'acides aminés qui sont enfuis ou exposés à l'état replié.

Cette méthode suppose qu'il existe une structure résiduelle à l'état déplié, où certaines positions sont plus enfuis que d'autres. En outre, cela doit rendre le paramétrage plus robuste et moins sensible à la taille et à la structure des homologues naturels utilisés pour définir les compositions d'acide aminés cibles, car les fréquences d'acide aminés des positions exposés et des régions enfouies sont calculées séparément. En principe, cela double le nombre d'énergie de référence à ajuster. Cependant, nous avons réduit ce nombre en introduisant des classes de similarité d'acide aminés, avec une énergie de référence ajustable par classe. Contrainte qui est levée en fin d'optimisation. Lors de l'optimisation de l'énergie de référence, nous effectuons, des calculs de séquences pour chaque protéines de notre jeu de test (à l'état apo) où une position sur deux peut muter, soit la moitié des positions (à l'exception de Gly et Pro), avec une simulations distinctes pour chaque moitié. De cette façon, lors de l'optimisation des paramètres, une position mutable est toujours entouré d'un environnement identique au type sauvage au moins sur les deux positions immédiatement voisines sur le squelette. Les calculs de conception des protéines s'appuient sur une méthode d'exploration Monte Carlo avec échange de réplique, puissante et efficace, qui utilise plus d'un demi-milliard de pas par simulation et par réplique, et produit des milliers de séquences dans une seule simulation. Les valeurs des énergie de référence sont optimisées avec deux choix différents du couple modèle de solvant constante diélectrique des protéines ϵ_p . *Les performances sont améliorées pour le modèle exact GB. Le modèle présente plusieurs limitations, dont la valeur de ϵ_p (78). Cette valeur diélectrique signifie que la structure protéique (y compris le squelette) est autorisée à être déformée. Cette valeur diélectrique a été développée dans Proteus, sur la base d'une méthode Monte Carlo hybride qui prend en compte les interactions électrostatiques et les interactions de van der Waals.*

Une autre limitation de notre modèle est la nécessité, pour des résultats optimaux, de paramétrer les énergies de référence spécifiquement pour un ensemble donné de protéines. Cette étape est bien automatisée et de façon très parallèle. Cependant, cela implique plusieurs choix qui sont partiellement arbitraires. Ceux-ci comprennent le choix d'un ensemble de domaines protéiques pour représenter la protéine ou la famille d'intérêt. Nous devons également choisir un seuil de similarité pour définir les homologues cibles à partir desquels sont calculées les compositions expérimentales d'acides aminés. Ici, nous

avons choisi d'utiliser les homologues de chaque membre de la famille, de calculer leurs compositions, puis la moyenne sur les six familles. Cette méthode a correctement fonctionné, mais d'autres choix sont possibles et il faut travailler d'avantage pour pouvoir tirer des conclusions définitives sur ces choix.

Une autre limitation de notre modèle est l'approximation de la position des chaînes latérales en rotamères discrets. qui nécessite une certaine adaptation de la fonction d'énergie pour éviter les affrontements stériques exagérés. La méthode utilisée ici est la méthode de minimisation de la paire de résidu décrite précédemment (26,76).

ailleurs !! Une cinquième limitation est l'utilisation d'un modèle de solvation additif par paires (comme dans la plupart des modèles CPD). Plus précisément, l'environnement diélectrique de chaque paire de résidus est supposé ici être celui de la structure native (ce qu'on appelle "Approvisionnement environnemental natif" ou NEA 74,76). Cela conduit à une fonction énergétique qui a la forme d'une somme sur les paires de résidus et peut être pré-calculée et stockée dans une matrice énergétique, qui sert alors de table de consultations pendant les simulations Monte-Carlo. Malgré cette approximation, le modèle a donné de bons résultats pour un grand nombre d'indices acide/base de référence, un problème très sensible au traitement électrostatique 74. Certaines de ces limitations ont été supprimées dans le modèle exactGB. En particulier, comme la fonction d'énergie est principalement basée sur la physique, elle a pu être améliorée par implémentations d'un calcul "GB" plus exact. Par ailleurs, il a été mis en place un modèle amélioré pour la solvation hydrophobe (80), qui est plus rapide et plus précise que notre terme d'énergie de surface actuelle (article en préparation).

4.8.2 modèle de test et applications

Les séquences calculées sont largement comparées aux séquences naturelles, à travers des tests de reconnaissance du pli, des calculs de similarité et des calculs d'entropie. Dans les simulations, nous concevons la totalité de la séquence de la protéine, de sorte que toutes les positions (à l'exception de Gly et Pro) peuvent muter librement, soumis seulement à un biais vers la composition moyenne expérimentale des acides aminés (à travers les énergies de références). Malgré la quasi-absence de biais expérimentaux ou de contraintes, les séquences obtenues ont une forte similitude globale avec les séquences naturelles de Pfam, mesurées par les scores de similarité Blosom40. Les scores obtenus sont, pour l'essentiel, comparables aux scores de similarité entre les paires de séquences Pfam entre-elles. Ainsi,

les séquences conçues avec Proteus ressemblent à des homologues naturels modérément éloignés. La similitude est très forte pour les résidus au coeur de la protéine, comme cela a été observé dans des études CPD précédantes (30,72). En revanche, pour les résidus à la surface de la protéine, les scores de similarité sont proche de zéro, le score obtenu si l'on tirait au hasard les acides aminés a ces positions. Notez que de nombreux résidus de surface sont impliqués dans des interactions fonctionnelles, comme les onze résidus de liaison aux peptides dans les domaines PDZ. Les résidus de surface sont également sélectionnés selon l'évolution pour éviter l'agrégation ou des adhésion indésirable. La plupart de ces contraintes fonctionnelles ne sont pas explicitement pris en compte dans notre protocole de design (bien que la fonction d'énergie soit indirectement la solubilité des protéines). Malgré les scores de similarité limités pour les régions de surface, la reconnaissance de pli avec l'outil Superfamily appliquée aux meilleures modèles concues est presque parfaite. Les test de reconnaissance de plie antérieurs qui utilisaient une fonction d'énergie plus simple donnaient un taux de reconnaissance de pli inférieur, à environ 85% (pour un ensemble de tests plus large et plus diversifié) et des similitudes inférieures (15,50). De toute évidence, l'utilisariion combinée d'un champ de force protéique amélioré, du solvant "GB" et des énergie de référence spécifiques à la famille conduisent à des séquences calculées sont proche des séquences natives et sans doute meilleures.

les séquences Proteus ont également été comparées aux séquences obtenues avec le logiciel Rosetta , qui a lui-même été testé de manière approfondie. Sur la Base des scores de similarité de Blosum (par rapport aux séquences naturelles dans Pfam) et des tests de reconnaissance du pli, les séquences Proteus et Rosetta semblent être de la même qualité. Cependant, rosetta fait moins de mutations que Proteus ; de sorte que les scores d'identité, par rapport à la protéine de type sauvage correspondante, sont environ 6% plus haut chez Rosetta. Ce qui veut dire que Proteus modifie environ 5 positions en plus , en moyenne, par domaine PDZ. Ce nombre peut facilement être réduit en ajoutant à la fonction d'énergie de Proteus un terme d'énergie de biais explicite qui augmente avec le nombre de mutations l'éloignant de la séquence de type sauvage.

Une autre possibilité pour tester nos séquences est d'utiliser les simulation de dynamique moléculaire. Nicolas Panel a testé dix séquences Tiam1 calculées dans des simulation MD assez longues, dans la forme apo, en utisant le même champ de force protéique que dans le modèle CPD (champ de force Amber) et un modèle d'eau explicite. Ces séquences

Conclusion

XXX

Annexe 1 :Liste des positions actives pour chaque test

Nom	S_{Vois}	positions actives
1A81 1	10	10 13 16 84 86
1A81 2	10	20 21 24 27 116
1A81 3	10	35 38 56 105 107
1A81 4	10	44 47 52 65 67
1A81 5	10	82 84 86 87 90
1ABO 1	10	64 66 90 93 100
1ABO 2	10	72 74 80 104 111
1ABO 3	10	79 82 102 111 115
1ABO 4	10	83 86 104 105 106
1ABO 5	10	93 100 102 113 116
1BM2 1	10	101 106 140 141 146
1BM2 2	10	120 128 131 132 135
1BM2 3	10	58 61 127 128 129
1BM2 4	10	74 75 98 100 105
1BM2 5	10	85 87 95 110 128
1CKA 1	10	136 138 158 175 190
1CKA 2	10	149 166 169 171 181
1CKA 3	10	151 153 157 159 172
1CKA 4	10	164 170 172 184 187
1CKA 5	10	172 174 182 186 187
1G9O 1	10	10 13 54 57 92
1G9O 2	10	15 39 42 54 57
1G9O 3	10	24 26 28 39 42
1G9O 4	10	48 53 57 59 88
1G9O 5	10	75 78 79 86 88
1M61 1	10	12 20 23 24 27
1M61 2	10	17 20 24 37 49
1M61 3	10	27 33 51 100 102
1M61 4	10	5 8 10 11 36
1M61 5	10	59 71 84 87 94
1O4C 1	10	20 21 32 34 46
1O4C 2	10	2 71 79 81 82
1O4C 3	10	33 45 63 71 73
1O4C 4	10	43 45 63 71 85
1O4C 5	10	8 33 82 83 86
1R6J 1	10	194 237 239 270 272
1R6J 2	10	199 201 211 218 232
1R6J 3	10	213 218 227 232 238
1R6J 4	10	221 227 232 267 269
1R6J 5	10	241 254 258 267 269
2BYG 1	10	189 191 221 244 246
2BYG 2	10	205 224 239 245 248
2BYG 3	10	232 233 265 272 274
2BYG 4	10	238 240 243 276 278
2BYG 5	10	253 261 264 265 274

Table 6 – Les tests avec cinq positions actives

Nom	S_{Vois}	positions actives
1A81 1	10	13 15 39 41 53 86 89 90 93 103
1A81 2	10	39 41 53 55 64 66 76 89 92 103
1A81 3	10	51 53 64 66 68 74 76 82 88 92
1A81 4	10	76 82 87 88 90 91 92 95 97 99
1A81 5	10	9 10 11 16 41 51 53 66 88 89
1ABO 1	10	64 72 74 79 89 91 101 103 108 111
1ABO 2	10	66 68 80 82 88 90 100 102 104 111
1ABO 3	10	69 70 72 74 80 81 106 113 114 115
1ABO 4	10	71 78 83 84 94 99 101 104 105 106
1ABO 5	10	72 79 82 94 99 102 104 106 111 115
1BM2 1	10	119 120 121 122 123 125 131 134 135 140
1BM2 2	10	125 126 127 129 130 133 134 136 137 147
1BM2 3	10	83 99 101 106 108 135 140 141 146 148
1BM2 4	10	85 95 97 110 118 120 125 128 131 132
1BM2 5	10	99 101 106 139 140 141 142 143 144 146
1CKA 1	10	134 135 160 161 162 173 174 175 176 179
1CKA 2	10	137 139 143 151 153 157 159 172 182 186
1CKA 3	10	138 140 147 149 150 155 166 169 181 188
1CKA 4	10	140 141 153 154 155 157 174 175 184 186
1CKA 5	10	151 153 157 166 168 173 174 176 178 179
1G9O 1	10	10 11 13 14 15 16 53 54 57 92
1G9O 2	10	15 17 24 26 39 42 48 51 53 88
1G9O 3	10	26 28 39 42 48 53 57 59 88 90
1G9O 4	10	34 35 58 60 68 70 74 75 89 91
1G9O 5	10	71 73 74 77 80 81 82 83 84 85
1M61 1	10	10 12 20 23 24 27 35 49 102 104
1M61 2	10	17 20 21 24 37 39 40 47 49 58
1M61 3	10	34 36 46 48 59 61 71 83 84 87
1M61 4	10	5 6 11 36 46 48 61 69 83 84
1M61 5	10	59 61 70 71 75 77 83 86 87 92
1O4C 1	10	31 33 45 47 61 63 73 86 89 100
1O4C 2	10	50 51 52 53 63 72 73 77 85 89
1O4C 3	10	61 62 63 71 72 73 79 85 88 89
1O4C 4	10	73 74 75 76 77 89 92 94 96 101
1O4C 5	10	90 91 93 96 98 99 101 102 103 104
1R6J 1	10	193 194 195 197 199 218 232 236 267 269
1R6J 2	10	199 209 211 213 218 227 232 238 265 267
1R6J 3	10	201 204 205 209 211 218 241 258 265 267
1R6J 4	10	209 211 213 218 227 238 241 258 265 267
1R6J 5	10	238 240 241 242 246 257 258 261 265 267
2BYG 1	10	194 196 203 205 224 233 239 245 274 276
2BYG 2	10	203 205 207 224 227 233 239 243 245 276
2BYG 3	10	206 207 222 245 248 253 256 261 264 265
2BYG 4	10	221 222 245 248 251 253 256 261 264 265
2BYG 5	10	247 248 249 250 251 252 259 262 263 275

Table 7 – Les tests avec dix positions actives

Conclusion

Nom	S_{Vois}	positions actives
1A81 1	1	9 11 12 13 15 16 17 19 20 25 28 39 40 41 42 43 44 45 114 117
1A81 2	1	9 11 12 13 15 16 17 19 20 25 28 39 40 41 42 43 44 45 47 117
1A81 3	1	9 11 12 13 15 16 17 19 19 41 43 48 51 68 74 84 86 109 114 117
1A81 4	1	12 13 15 16 17 19 20 25 28 39 40 41 42 43 44 45 47 86 114 117
1A81 5	1	13 15 16 19 41 43 48 51 60 64 68 70 71 74 84 86 87 88 109 114 117
1ABO 1	1	64 66 67 68 82 86 87 88 89 90 91 101 102 102 103 103 108 111 113 116
1ABO 2	1	64 65 65 66 67 84 87 88 89 90 91 93 100 101 102 103 108 111 113 116
1ABO 3	1	65 66 67 87 88 89 90 91 93 94 95 100 101 102 103 106 108 111 113 116
1ABO 4	1	64 65 66 67 69 87 88 89 90 91 93 100 101 102 103 106 108 111 113 116
1ABO 5	1	66 67 68 82 86 87 88 89 90 91 93 100 101 102 103 106 108 111 113 116
1BM2 1	1	55 56 60 61 62 83 84 85 86 87 95 97 99 110 118 125 127 133 150 152
1BM2 2	1	55 56 60 61 62 83 84 85 86 87 95 97 99 110 118 125 127 128 129 152
1BM2 3	1	55 56 58 60 61 62 64 67 69 73 83 84 85 86 87 129 132 133 150 152
1BM2 4	1	55 56 60 61 62 69 83 84 85 86 87 95 97 99 110 129 132 133 150 152
1BM2 5	1	58 60 60 61 61 62 64 67 69 73 75 83 84 85 86 129 132 133 150 152
1CKA 1	1	134 135 136 137 138 139 150 151 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 2	1	134 135 136 137 139 150 151 153 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 3	1	134 136 137 139 150 151 157 158 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 4	1	136 137 139 150 151 153 158 159 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 5	1	137 139 150 151 153 158 160 161 162 163 164 170 171 172 173 174 175 179 189 190
1G9O 1	1	9 10 11 13 14 15 31 34 38 54 57 58 60 68 90 91 92 94 95 96
1G9O 2	1	9 11 13 14 15 16 31 34 38 54 57 58 60 68 90 91 92 94 95 96
1G9O 3	1	9 11 13 14 15 31 34 38 54 55 57 58 60 68 90 91 92 94 95 96
1G9O 4	1	9 11 13 15 16 17 54 57 58 59 60 61 68 89 90 91 92 94 95 96
1G9O 5	1	10 11 13 15 16 17 54 57 58 60 61 68 89 90 90 91 92 94 95 96
1M61 1	1	34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82
1M61 2	1	35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83
1M61 3	1	38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85 87
1M61 4	1	42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85 87 88 98
1M61 5	1	5 7 50 61 63 69 71 77 78 81 82 83 84 85 87 88 98 103 104 109
1O4C 1	1	32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 84 85 86 87 89
1O4C 2	1	3 4 5 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79
1O4C 3	1	1 3 4 5 6 7 8 9 11 17 31 32 33 35 43 45 65 81 82 83
1O4C 4	1	1 2 3 4 5 6 7 8 9 11 12 13 14 17 19 35 65 81 82 83
1O4C 5	1	1 3 4 5 6 7 8 9 11 12 17 31 32 33 34 35 65 81 82 83
1R6J 1	1	193 194 195 197 214 215 217 218 233 235 236 237 239 240 241 242 247 269 270 273
1R6J 2	1	193 194 197 198 199 217 233 235 236 237 238 239 240 241 242 247 268 270 272 273
1R6J 3	1	193 195 197 217 233 235 236 239 240 241 242 244 245 247 268 269 270 270 272 273
1R6J 4	1	193 195 197 217 233 235 236 237 239 241 242 244 245 247 268 269 270 272 273 273
1R6J 5	1	193 194 197 198 199 233 236 237 239 239 240 241 247 268 268 269 270 270 272 273
2BYG 1	1	186 187 188 189 190 191 192 215 216 219 244 246 270 271 273 274 278 280 281 282
2BYG 2	1	186 187 188 189 190 215 216 219 221 223 240 243 270 271 273 274 278 280 281 282
2BYG 3	1	186 187 188 189 190 215 216 219 221 223 240 243 244 270 271 273 278 280 281 282
2BYG 4	1	186 187 188 189 190 215 216 219 221 223 240 243 244 270 274 276 278 280 281 282
2BYG 5	1	187 189 190 191 192 215 216 219 221 223 240 243 244 270 274 276 278 280 281 282

Table 8 – Les tests avec vingt positions actives

Nom	S_{Vois}	positions actives
1A81 1	1	9 11 12 13 15 16 17 19 20 25 26 27 28 29 36 38 39 40 41 42 43 48 51 68 74 84 86 109 114 117
1A81 2	1	9 10 11 12 13 15 16 17 19 20 25 28 39 41 43 48 51 68 74 83 84 86 87 88 90 91 93 109 114 117
1A81 3	1	9 11 12 13 15 16 17 19 20 25 27 28 36 38 39 40 41 41 42 43 43 48 51 68 74 84 86 109 114 117
1A81 4	1	9 10 11 12 13 15 16 17 19 20 25 28 36 39 40 41 42 43 43 44 45 48 51 68 74 84 86 109 114 117
1A81 5	1	9 10 11 12 13 15 16 17 19 20 25 28 39 40 41 42 43 44 45 47 48 51 52 68 74 84 86 109 114 117
1ABO 1	1	64 65 66 67 68 70 71 72 75 78 79 80 81 82 83 86 87 88 89 90 91 93 100 101 102 103 108 111 113 116
1ABO 2	1	64 65 66 67 68 72 75 78 80 81 82 83 84 86 87 88 89 90 91 93 94 100 101 102 103 104 108 111 113 116
1ABO 3	1	64 66 67 68 70 71 72 78 82 86 87 88 89 90 91 93 94 95 96 99 100 101 102 103 104 105 108 111 113 116
1ABO 4	1	64 65 66 67 70 71 72 68 82 86 87 88 89 90 91 93 94 95 96 99 100 101 102 103 104 105 108 111 113 116
1ABO 5	1	65 66 67 70 71 72 75 78 80 82 86 87 88 89 90 91 93 94 95 96 99 100 101 102 103 108 111 113 116
1BM2 1	1	55 56 58 60 61 62 83 84 85 86 87 95 97 99 110 118 119 120 121 122 125 127 128 129 130 131 132 133 150 152
1BM2 2	1	56 58 60 61 62 83 84 85 86 87 95 97 99 110 118 119 120 121 122 123 125 127 128 129 130 131 132 133 150 152
1BM2 3	1	58 60 61 62 83 84 85 86 87 95 97 99 108 109 110 118 120 121 122 123 125 127 128 129 132 133 134 135 150 152
1BM2 4	1	55 56 58 60 61 62 83 84 85 86 87 95 97 99 108 109 110 118 120 121 125 127 128 129 130 131 132 133 150 152
1BM2 5	1	56 58 60 61 62 67 83 84 85 86 87 95 97 99 110 111 112 113 115 118 125 127 128 129 130 131 132 133 150 152
1CKA 1	1	134 135 136 137 139 140 141 142 143 144 146 147 148 149 150 151 158 159 160 161 162 163 164 170 171 172 173 179 189 190
1CKA 2	1	134 135 136 137 139 143 144 146 147 148 149 150 151 158 159 160 161 162 163 164 170 171 172 173 179 186 187 188 189 190
1CKA 3	1	135 136 137 139 144 146 147 148 149 150 151 157 158 159 160 161 162 163 163 164 170 171 172 173 179 186 187 188 189 190
1CKA 4	1	136 137 139 140 141 142 143 144 146 147 148 151 158 159 160 161 162 163 164 170 171 172 173 179 184 186 187 188 189 190
1CKA 5	1	134 136 137 139 140 141 142 143 144 146 147 148 151 158 159 160 161 162 163 164 170 171 172 173 179 182 187 188 189 190
1G90 1	1	9 10 11 13 15 24 31 34 38 40 41 42 43 46 48 49 50 51 54 57 58 60 68 89 90 91 92 94 95 96
1G90 2	1	9 11 13 15 31 32 34 38 40 41 42 43 46 48 49 50 51 54 57 58 60 68 89 90 90 91 92 94 95 96
1G90 3	1	9 10 11 13 15 31 32 34 38 40 41 42 43 46 48 49 50 51 54 57 58 60 68 89 90 91 92 94 95 96
1G90 4	1	10 11 13 14 15 31 32 34 38 40 41 42 43 46 48 49 50 51 54 57 58 60 61 68 89 90 91 92 94 95 96
1G90 5	1	10 11 13 14 15 31 32 40 41 42 43 46 48 49 50 51 54 57 58 60 61 62 68 87 89 90 91 92 94 95 96
1M61 1	1	12 14 15 20 34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85 87 88 98
1M61 2	1	6 7 8 10 11 12 14 15 20 21 34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82
1M61 3	1	5 7 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85 87 88 98 103 104 109
1M61 4	1	7 8 10 11 12 14 15 20 34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84
1M61 5	1	8 10 11 12 14 15 20 34 35 36 37 38 39 42 46 47 48 49 50 61 63 69 71 77 78 81 82 83 84 85
1O4C 1	1	1 2 3 4 5 6 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 81 82 83 90 91 92 93 94 96
1O4C 2	1	1 3 4 5 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 90 91 92 93 96
1O4C 3	1	1 3 4 5 6 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 90 91 92 93 96
1O4C 4	1	1 3 4 5 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 84 91 92 93 96
1O4C 5	1	1 3 4 5 6 7 8 9 11 17 31 32 33 34 35 43 45 63 65 71 73 79 80 81 82 83 84 92 93 96
1R6J 1	1	193 194 195 197 198 199 217 218 219 220 221 222 223 224 225 226 227 228 229 233 235 236 237 239 247 268 269 270 272 273
1R6J 2	1	193 194 195 197 198 199 217 220 221 222 223 224 225 226 227 228 229 233 235 235 236 237 239 240 247 268 269 270 272 273
1R6J 3	1	193 194 195 197 198 199 208 217 220 221 222 223 224 225 226 227 228 229 230 233 235 236 237 239 247 268 269 270 272 273
1R6J 4	1	193 194 195 197 198 199 217 220 221 222 223 224 225 226 227 228 229 233 235 236 237 239 240 247 249 268 269 270 272 273
1R6J 5	1	194 195 197 198 199 217 220 221 222 223 224 225 226 227 228 229 233 235 236 237 239 240 247 249 250 268 269 270 272 273
2BYG 1	1	186 187 188 189 190 191 192 215 216 219 221 223 240 243 244 246 250 251 252 253 254 255 256 257 259 260 278 280 281 282
2BYG 2	1	186 187 188 189 190 191 192 215 216 219 221 223 240 243 244 246 250 251 252 253 254 255 256 257 259 260 278 280 281 282
2BYG 3	1	186 187 188 189 190 190 191 192 215 216 219 221 223 240 243 244 246 251 252 253 254 255 256 257 259 260 278 280 281 282
2BYG 4	1	186 187 188 189 190 191 192 215 216 219 221 223 240 243 244 245 246 251 252 253 254 255 256 257 259 260 278 280 281 282
2BYG 5	1	186 187 188 189 190 191 192 215 216 219 221 223 240 243 244 251 252 253 254 255 256 257 259 260 278 280 281 282

Table 9 – Les tests avec trente positions actives

Résum

Titre de la thèse

XXX

Mots-clés : motclé1, motclé2, motclé3

Abstract

Thesis title

XXX

Keywords: keyword1, keyword2, keyword3