Étude computationnelle du domaine PDZ de Tiam1

Nicolas Panel

NNT : 2017SACLX062

Thèse de doctorat de l'Université Paris-Saclay préparée à l'École Polytechnique

École doctorale n°573
INTERFACES : approches interdisciplinaires : fondements, applications et innovation
Spécialité de doctorat : Biologie

par

M. Nicolas Panel

Étude computationnelle du domaine PDZ de Tiam1

Thèse présentée et soutenue à l'École Polytechnique, le 7 novembre 2017.

Composition du Jury :

Mme.	Nathalie Reuter	Professeur Université de Bergen	(Rapporteur)
M.	Serge Crouzy	Directeur de recherche CEA - UMR 5249	(Rapporteur)
M.	Alexandre de Brevern	Directeur de recherche INSERM - UMR 1134	(Président)
Mme	Sophie Sacquin-Mora	Chargée de Recherche CNRS - UPR 9080	(Examinatrice)
Mme.	Jessica Andreani	Chargée de Recherche CEA - UMR 9198	(Examinatrice)
M.	Thomas Simonson	Professeur École Polytechnique - UMR 7654	(Directeur de thèse)

Remerciements

Je tiens tout d'abord à remercier les rapporteurs *Nathalie Reuter* et *Serge Crouzy* ainsi que les examinateurs *Jessica Andreani*, *Sophie Sacquin-Mora* et *Alexandre De Brevern* pour le temps consacré à l'évaluation de ce travail de thèse. Je remercie également *Yves Méchulam*, directeur du laboratoire de Biochimie, de m'avoir accueilli au sein du laboratoire.

Je tiens tout particulièrement à remercier *Thomas Simonson* qui a encadré mes recherches durant ces trois années. Je le remercie pour la confiance qu'il m'a accordée et sa disponibilité tout au long de ma thèse. Ses conseils, ses suggestions et sa rigueur scientifique ont grandement participé à la cohérence de ce travail. J'ai également appris énormément à travers nos discussions. Enfin, je le remercie pour ses relectures de ce manuscrit qui ont contribué à le rendre plus clair et plus fluide.

Je remercie chaleureusement l'ensemble des collègues de BIOC qu'ils soient encore présents ou partis vers de nouveaux horizons. Je remercie tout particulièrement *Karen*, *Clara* et *Pierre-Damien* pour ces trois ans passés ensemble, pour nos soirées apéro et nos quelques escapades à l'autre bout de l'Europe. Je souhaite également remercier *Francesco* et *Vaitea* pour les agréables moments passés au labo, autour d'un verre ou au bord du lac à déguster du vin tahitien. Je remercie *Thomas G.* de m'avoir fait découvrir le monde de la bioinformatique structurale lors de mon stage de M1. Je remercie également *David* et *Xingyu* pour notre collaboration dans l'application du CPD aux domaines PDZ. Je remercie *Pierre* et *Marc D.* pour les déjeuners passés ensemble. Merci *Michel* pour nos discussions musicales lors de tes passages au laboratoire. Je tiens aussi à remercier *Bon-Min* pour ses passages furtifs dans le bureau et les soirées passées ensemble (il nous reste encore quelques extensions de Carcassonne à tester !). Enfin, un grand merci à *Alexey, Zoltán, Savvas, Michou, Titine, Mimi, Emma, Marc G., Nhan, Amlam, Ditipriya, Giuliano, Nathalie, Guillaume, Mélanie* et *Caroline* qui ont, chacun à sa façon, participé à l'excellente ambiance de ce laboratoire.

Je tiens également à remercier nos collaborateurs de l'université de l'Iowa *Ernesto J. Fuentes*, *Young Joo Sun*, *Michael J. Schnieders* et *Jacob Litman* dont les discussions ont grandement profité aux orientations prises au cours de mes recherches et sans qui les validations expérimentales n'auraient pas été possibles. Je remercie GENCI pour les heures de calculs allouées au CINES et sur le TGCC.

Pour terminer, je tiens à remercier profondément tous les membres de ma famille ainsi que mes proches. Je remercie tout d'abord ma *Maman* sans qui je ne serais probablement jamais arrivé jusqu'ici. Merci d'avoir toujours cru en moi et de n'avoir jamais douté des choix je prenais au cours de ces (très) longues années d'étude. Enfin merci d'avoir eu le courage de lire (et relire !) ce manuscrit qui a pu parfois te paraitre bien obscure. Je tiens également à remercier mes frère et sœurs, *Mathieu*, *Marion* et *Camille* qui ont toujours été présents. Je tiens bien évidement à remercier *Ariane* pour tout le soutien que tu m'as apporté pendant ces trois années. Merci d'avoir toujours été à mon écoute et de m'avoir encouragé. Et surtout un grand merci pour être prête à me suivre à l'autre bout du monde (au moins de l'Europe) ! Enfin, je souhaiterais remercier *Benoit* et *Morgan* pour nos débuts de weekend à la Butte aux cailles qui m'ont permis de relâcher la pression à certains moments.

Merci à tous...

Table des matières

Liste des figures

Liste des tableaux

Première partie

Introduction

Chapitre 1

Interactions protéine-protéine et domaines PDZ

Dans la cellule, les interactions protéine-protéine (PPI) jouent un rôle majeur dans des activités aussi variées que la transduction du signal, le contrôle de l'activité enzymatique ou la conversion d'énergie en mouvement. Toutes ces activités doivent être finement régulées pour le bon fonctionnement de la cellule, leur dérégulation étant souvent responsable de pathologies comme les cancers. L'intérieur de la cellule est un milieu dense dans lequel les protéines sont en constante interaction les unes avec les autres et doivent donc pouvoir reconnaitre spécifiquement leurs partenaires. Cette reconnaissance est assurée par des domaines protéiques capables de lier une région située à la surface du partenaire et possédant des caractéristiques physico-chimiques (charge, hydrophobicité, etc...) particulières. Il existe différents types de domaines de reconnaissance, chacun spécialisé dans la reconnaissance d'un type d'interface. Par exemple, les domaines SH3 se lient à des régions riches en prolines quand les domaines PDZ reconnaissent majoritairement l'extrémité C-terminale du partenaire. D'autres domaines comme les domaines SH2, PTB et FHA se fixent sur des régions modifiées des protéines (phosphorylation, acétylation et méthylation).

Au cours de cette étude nous nous intéresserons à la famille des domaines PDZ et plus particulièrement au domaine PDZ de la protéine Tiam1 qui est impliquée dans l'organisation et la motilité cellulaire. Nous présenterons dans un premier temps le rôle et l'organisation des domaines PDZ au sein de la cellule ainsi que les études computationnelles ayant été effectuées sur ces domaines. Nous détaillerons ensuite le rôle et les données disponibles concernant la protéine Tiam1 et son domaine PDZ.

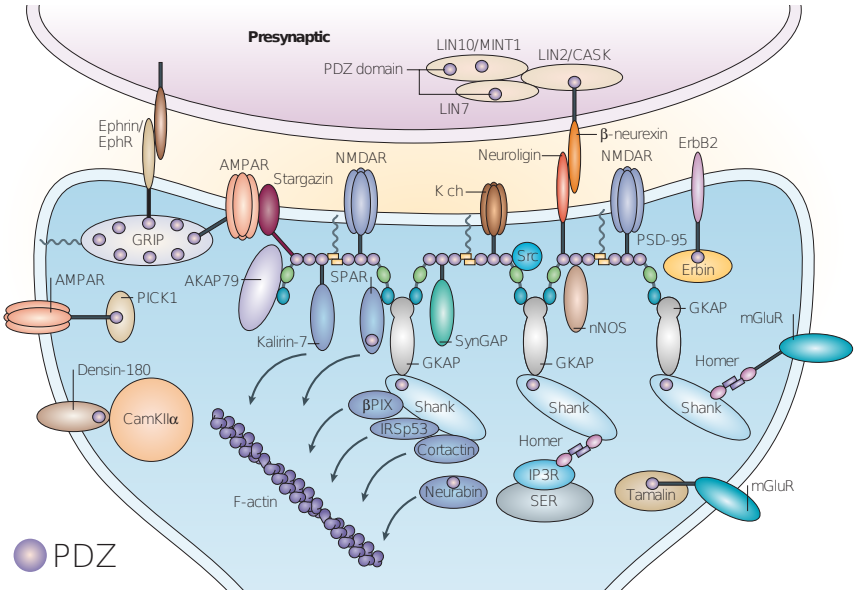


Figure 1.1 – Représentation schématique de la répartition des domaines PDZ dans une synapse excitatrice de mammifère. Les domaines PDZ sont représentés par les ronds violets. Seul un sous-ensemble de protéines connues est représenté. D'après Kim & Sheng [2004].

1.1 Les domaines PDZ

Les domaines PDZ (*Postsynaptic protein-95/Disk large/Zonula occludens-1*) font partie des domaines de reconnaissance les plus abondants dans le règne du vivant. On estime actuellement à environ 270 le nombre de domaines présents dans le génome humain, répartis sur 150 protéines (Luck *et al.* [2012]). Certaines protéines, comme MUUP1, possèdent jusqu'à 13 domaines PDZ (Ye & Zhang [2013]). La présence de ces domaines au sein des protéines permet de reconnaître et de lier spécifiquement les partenaires protéiques (Baruch & Wendell [2001]). Les domaines PDZ sont impliqués dans un grand nombre de processus cellulaires comme la polarité, la migration, la croissance et la différenciation cellulaire ou encore le développement embryonnaire. La figure 1.1 illustre par exemple leur rôle dans l'organisation d'une synapse. On constate que la majorité des interactions font intervenir un ou plusieurs domaines PDZ.

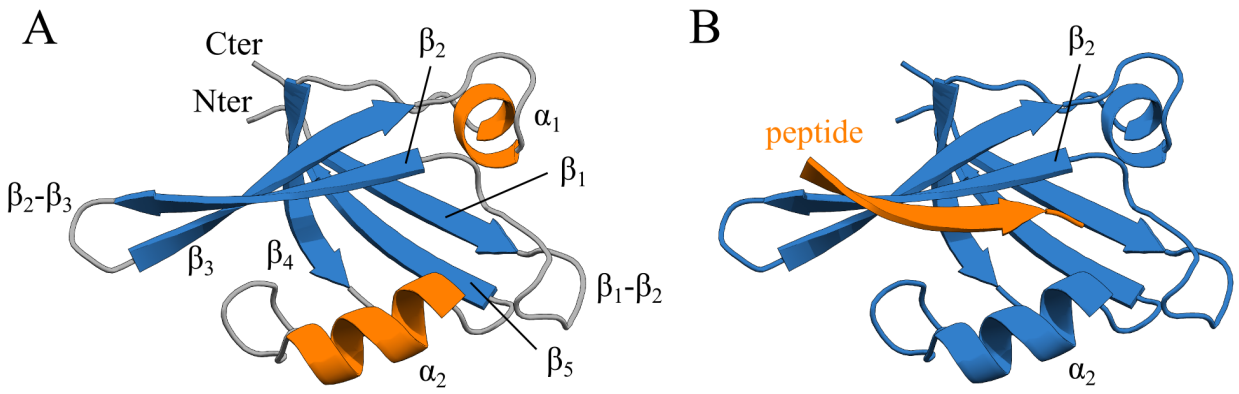


Figure 1.2 – Organisation des domaines PDZ. A : Organisation des structures secondaires d'un domaine PDZ. Les hélices α sont en orange, les brins β en bleu et les boucles en gris. B : Complexe formé par un domaine PDZ (bleu) et l'extrémité C-terminale de son partenaire (orange).

1.1.1 Organisation des domaines PDZ

1.1.1.1 Structure des domaines PDZ

Les PDZ sont des petits domaines globulaires d'environ 90 acides aminés. Ils présentent tous une structuration similaire composée de cinq ou six brins β (β_1 à β_6) et deux hélices α (α_1 et α_2), reliées par des boucles de tailles variables (figure 1.2A). D'autres structures secondaires sont parfois présentes au niveau des extrémités N- et C-terminales, qui modulent la structure et la fonction du domaine (Wang *et al.* [2010]). Dans une protéine, les domaines PDZ peuvent être seuls ou organisés en tandem, chaque domaine étant alors séparé du suivant par une séquence courte et très conservée. Dans cette organisation, les domaines PDZ ne fonctionnement pas indépendamment les uns des autres mais s'associent pour former des supramodules (Ye & Zhang [2013]). L'activité et la stabilité des domaines sont alors dépendantes des autres PDZ (Long *et al.* [2008]).

1.1.1.2 Mécanismes de reconnaissance

Les domaines PDZ lient généralement l'extrémité C-terminale de leur partenaire qui vient former un brin β au niveau d'un sillon situé entre β_2 et α_2 (figure 1.2B). D'autres modes de liaison ont cependant été observés, notamment avec des boucles internes (Gee *et al.* [1998] ; Hillier [1999] ; Wong *et al.* [2003]), des régions intrinsèquement désordonnées (Ivarsson [2012])

et des phospholipides (Gallardo *et al.* [2010]). Nous nous intéressons ici uniquement au mode de reconnaissance le plus courant, c'est-à-dire la liaison d'une extrémité C-terminale.

Les domaines PDZ reconnaissent les quatre à sept acides aminés C-terminaux de leurs partenaires (Appleton *et al.* [2006] ; Tonikian *et al.* [2008]) mais peuvent également lier des peptides d'une dizaine d'acides aminés. Par convention, ces positions sont numérotées négativement, la position P_0 correspondant à l'extrémité C-terminale. La nature des acides aminés présents à l'interface de liaison est responsable de la grande spécificité des domaines PDZ pour leurs partenaires. Cette spécificité ne semble cependant pas dépendre exclusivement de la nature des résidus de l'interface puisque deux domaines PDZ liant le même ligand ne présentent pas obligatoirement la même séquence (Ernst *et al.* [2010]). Certaines zones d'ombres persistent donc sur les mécanismes mis en jeu dans l'interaction des PDZ avec leurs partenaires. L'analyses des structures cristallographiques et l'étude de bibliothèques combinatoires de peptides a cependant permis d'identifier certaines régions importantes dans la liaison de l'extrémité C-terminale de la protéine cible. Ainsi, la boucle β_1-β_2 est responsable de la liaison aspécifique du partenaire tandis que les résidus présents au niveau des brins β_2 et β_3 et de l'hélice α_1 jouent un rôle dans la spécificité de la liaison. L'interface entre le domaine PDZ et l'extrémité C-terminale est présentée en figure 1.3. La boucle β_1-β_2 présente le motif R/K-XXX-G-Φ-G-Φ (où Φ est un acide aminé hydrophobe) qui est très conservé au sein des domaines PDZ. Cette région est impliquée dans la liaison du groupement carboxylate de la position P_0 (Doyle *et al.* [1996] ; Songyang *et al.* [1997]). Cette interaction s'effectue avec le squelette de la deuxième glycine et la chaine latérale du résidu R/K, par l'intermédiaire d'une molécule d'eau.

Les régions responsables de la spécificité d'un domaine pour sa cible sont l'hélice α_2 et le brin β_2, qui forment l'interface de liaison, mais également le brin β_3. La position P_0 est certainement la plus importante dans la reconnaissance du partenaire. Cette dernière est la plupart du temps enfouie au sein d'une poche hydrophobe formée par la boucle β_1-β_2, α_2 et β_2 appelée S_0. Une seconde sphère de résidus autour de S_0 est également importante pour la spécificité (Murciano-Calles *et al.* [2014]). La poche S_0 reconnait préférentiellement les résidus de type hydrophobe mais est souvent optimisée pour un type particulier (Tonikian *et al.* [2008]). La position P_{-1} est le plus souvent un résidu hydrophobe mais peut, dans certains cas, être un résidu chargé (Ernst *et al.* [2014]). La spécificité à cette position provient des résidus présents au niveau des feuillets β_2 et β_3 et d'interactions avec le squelette de β_2. La

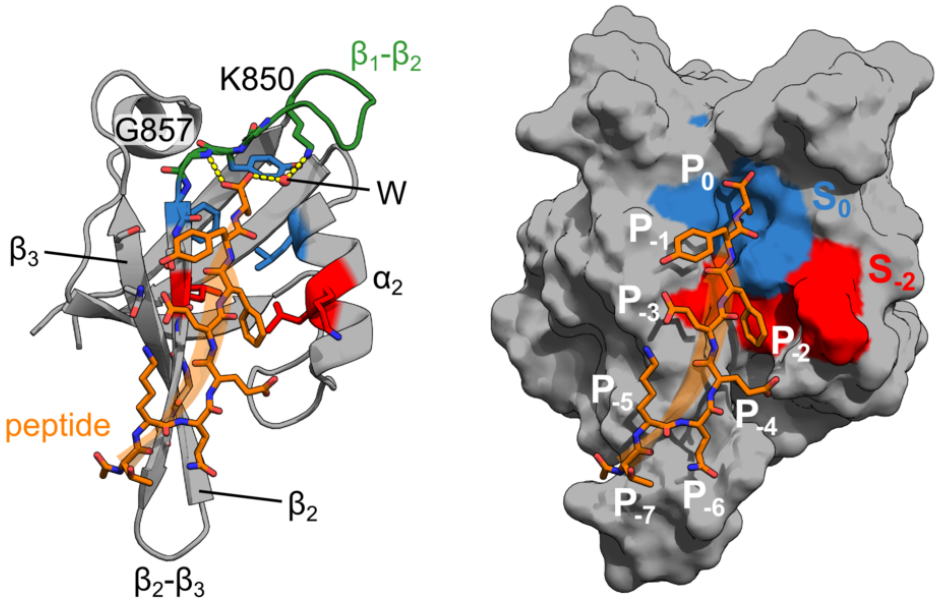


Figure 1.3 – Interface de liaison PDZ:peptide. Le domaine PDZ de Tiam1 lié au peptide Sdc1 est représenté en *cartoon* (gauche) et par sa surface (droite). Les résidus de la poche S_0 sont en bleu, ceux de la poche S_{-2} sont en rouge. La boucle β_1-β_2 présentant le motif R/K-XXX-G-Φ-G-Φ (où Φ est un acide aminé hydrophobe) est en vert.

position P_{-2} est également une position importante dans la spécificité. Elle est reconnue au niveau d'une poche appelée S_{-2} composée de résidus de β_2 et α_2. La position P_{-3} interagit avec des résidus de β_2 et β_3. On trouve souvent un résidu hydrophobe à cette position bien que dans certains cas, les types E/D et S/T soient préférés (Ernst *et al.* [2014]). Au-delà de la position P_{-3}, le rôle des résidus sur la reconnaissance semble très dépendante du système.

1.1.1.3 Classification des domaines PDZ

Selon la nature des résidus reconnus aux positions P_0 à P_{-2}, les domaines PDZ ont été classés en trois groupes (Songyang *et al.* [1997] ; Stricker *et al.* [1997] ; Nourry *et al.* [2003]) : S/T-X-Φ pour la classe I, Φ-X-Φ pour la classe II et D/E-X-Φ pour la classe III (où Φ est un acide aminé hydrophobe). Une quatrième classe correspondant au motif X-Ψ-D/E (où Ψ correspond à une acide aminé aromatique) est parfois ajoutée (Vaccaro & Dente [2002]). Cette classification ne prend en compte que les trois résidus C-terminaux et est donc imparfaite puisque les résidus en amont de la position P_{-2} jouent également un rôle dans la spécificité. En analysant environ 3100 peptides reconnus par des domaines PDZ, Tonikian *et al.* [2008] ont défini 16 classes basées sur les 6 positions C-terminales. La classification des domaines

PDZ reste cependant toute relative car de nombreux chevauchements sont observés entre les différentes classes (Stiffler *et al.* [2007]).

1.1.2 Les domaines PDZ comme cibles thérapeutiques

De par leur implication dans un grand nombre de processus cellulaires, les domaines PDZ sont d'excellentes cibles thérapeutiques dans le traitement de certaines pathologies, notamment neuronales, ou dans les cancers. En effet, en inhibant un domaine PDZ, il est possible d'interrompre une voie de signalisation. Différents domaines PDZ se sont révélés être de bonnes cibles comme PSD-95 (Aarts [2002]) et PICK1 (Thorsen *et al.* [2009]) dans le cas de maladies neuronales, et NHERF1 (Mayasundari *et al.* [2008]), AF6 (Joshi *et al.* [2006]) et MAGI3 (Fujii *et al.* [2003]) dans le cas de cancers.

Le développement de petites molécules inhibitrices des PPI est difficile en raison de la surface de contact importante entre les deux partenaires (Wells & McClendon [2007]) qui, dans le cas des domaines PDZ, est un long et profond sillon (Fry & Vassilev [2005]). Une alternative consiste à développer des peptides inhibiteurs en se basant sur la séquence des partenaires naturels. Cette approche a notamment été testée sur le domaine PDZ2 de la protéine PSD-95 (Aarts [2002]) et la protéine CAL (Amacher *et al.* [2014]). Dans certains cas, il est également possible de cibler deux domaines PDZ organisés en tandem en reliant deux peptides par un lien. Cette approche a été appliquée avec succès aux domaines PDZ1-2 de PSD-95 et augmente l'affinité de l'inhibiteur (Bach *et al.* [2012]). L'une des principales difficultés à l'utilisation de peptides thérapeutiques est leur instabilité *in vivo*. Elle peut toutefois être réduite par l'incorporation d'acides aminés non naturels (Udugamasooriya *et al.* [2008] ; Patra *et al.* [2012]), la modification chimique du squelette (Bach *et al.* [2011]) ou l'utilisation de peptides cycliques (LeBlanc *et al.* [2010]).

1.1.3 Études computationnelles existantes

L'étude des domaines PDZ par des approches expérimentales pour identifier de nouveaux partenaires, caractériser la nature des interactions au sein des complexes ou créer des inhibiteurs à des fins thérapeutiques, est un processus long est couteux. L'outil informatique permet dans certains cas d'accéder à ces informations de manière plus rapide. Nous détaillons dans cette partie quelques exemples d'études computationnelles des domaines PDZ.

1.1.3.1 Caractérisation des domaines PDZ par dynamique moléculaire

Les approches de simulation sont d'excellents outils pour caractériser la dynamique des domaines PDZ et identifier les résidus impliqués dans la reconnaissance des partenaires. En couplant l'utilisation d'un réseau élastique avec la théorie des graphes, Raimondi *et al.* [2013] ont ainsi pu observer une modification des mouvements corrélés des résidus au sein du domaine PDZ de la protéine PTP1E suite à la fixation d'un peptide. Ces modifications ont ensuite été validées par des expériences de RMN qui montrent un bon accord avec les prédictions. Également dans l'optique de prédire les changements de dynamique liés à la liaison d'un peptide, Cilia *et al.* [2012] ont développé une méthode mesurant les changements de dynamique des chaines latérales du domaine PDZ suite à la liaison d'un peptide. Cette approche, basée sur un échantillonnage Monte Carlo des chaines latérales a été appliquée pour prédire la dynamique des groupements méthyles des variants humain et murin du domaine PDZ2 de PTP1e.

Différentes études ont permis de caractériser les mécanismes impliqués dans la formation de complexes PDZ:peptide. Sensoy & Weinstein [2015] ont effectué des simulations de dynamique moléculaire du domaine PDZ de GIPC1 en présence de la structure complète de son partenaire Hx-8, inséré dans une membrane. Lors des simulations, l'extrémité C-terminale de Hx-8 vient spontanément interagir avec la boucle β_1-β_2, confirmant l'importance du motif conservé G-Φ-G-Φ. Blöchliger *et al.* [2015] se sont intéressés aux mécanismes impliqués dans l'association et la dissociation de l'extrémité C-terminale (correspondant à un peptide de six résidus) de la protéine RAGEF2 au domaine PDZ2 de PTP1E. Pour cela, dix simulations indépendantes des formes dissociée et associée ont été produites, totalisant 57 μs. Dans la moitié des simulations d'association, le peptide se lie au domaine PDZ dans une conformation proche de la conformation expérimentale. Dans trois simulations de dissociation, le peptide se réassocie à la protéine après dissociation complète. Cette étude a montré que le processus d'association et de dissociation du peptide faisait intervenir des interactions transitoires entre le groupe carboxylate du C-terminal et des résidus chargés positivement situés sur le pourtour du site de liaison. Elle a également montré que la boucle β_1-β_2 et la poche P_0 sont importants et favorisent la liaison du peptide.

L'étude des complexes PDZ:peptide par simulation de dynamique moléculaire permet également d'identifier les interactions responsables de la spécificité des domaines et de déterminer

la nature des interactions entre les deux partenaires. Dans une étude portant sur l'analyse de douze complexes PDZ par simulation, Basdevant *et al.* [2006] ont ainsi identifié les mécanismes moléculaires et thermodynamiques responsables de la spécificité de ces domaines. À partir des simulations, la contribution à l'énergie libre de liaison des énergies électrostatiques et apolaires et de l'entropie conformationnelle a été évaluée par MM/PBSA combiné avec une analyse quasi-harmonique. Cette étude a prédit que la composante apolaire représentait 77% de l'énergie libre absolue de liaison. Il a également été montré que la dynamique et l'entropie configurationnelle étaient différentes selon les domaines PDZ, ce qui peut jouer un rôle dans la reconnaissance des partenaires. Dans un second temps, l'énergie libre associée à la réorganisation des peptides et des protéines a été évaluée, suggérant que la réorganisation du peptide constituait une phase critique dans le processus de liaison. Dans une étude portant sur le domaine PDZ3 de PSD-95 par des simulations QM/MM couplées à des analyses MM/PBSA, Tian *et al.* [2011], ont montré que les positions P_0 et P_{-2} jouaient un rôle principal dans la liaison du partenaire et que les positions P_{-1} et P_{-3} participaient à la stabilité du complexe. Mamonova *et al.* [2017] ont mis en évidence qu'en plus de la position P_{-3}, les positions P_{-5} et P_{-6} étaient également impliquées dans la stabilité des complexes PDZ:peptide.

1.1.3.2 Prédiction des séquences liantes et des affinités

La prédiction des motifs reconnus par un domaine PDZ peut être utile pour l'identification de partenaires potentiels ou la création de peptides inhibiteurs. Différentes méthodes expérimentales permettent d'identifier à grande échelle les partenaires d'une protéine comme les micropuces et le *phage display*. Ces méthodes sont cependant longues à mettre en place et couteuses. À partir des données déjà existantes sur les domaines PDZ (micropuces, expériences de *phage display*, structures, banques de données d'interactions), plusieurs modèles basés sur la méthode des machines à vecteurs de support (SVM) ont été développés (Hui *et al.* [2013] ; Kundu & Backofen [2014] ; Nakariyakul *et al.* [2014]). Ces méthodes semblent donner de bons résultats avec notamment des valeurs d'aire sous la courbe ROC comprises généralement entre 0,80 et 0,90 en validation croisée.

Une autre méthode de prédiction utilisant la structure des complexes a été proposée par Smith & Kortemme [2010]. Contrairement aux SVM, cette méthode se base uniquement sur la structure tridimensionnelle du complexe d'intérêt pour identifier les séquences du peptide compatibles avec le domaine PDZ. La méthode génère tout d'abord un ensemble de confor-

mations à l'aide du programme Rosetta backrub (Lauck *et al.* [2010]). Dans un second temps, 2000 séquences du peptide sont générées par conformation à l'aide du programme Rosetta qui utilise un algorithme d'exploration de type Monte Carlo. À chaque séquence est associée un score. Les séquences générées sont ensuite rassemblées dans une matrice et pondérées par leur score respectif. La matrice obtenue représente alors le profil des séquences reconnues par le domaine. Cette méthode a été testée sur 17 domaines PDZ différents et donne des résultats en accord avec les expériences de *phage display* avec notamment 70 à 80% des acides aminés les plus fréquents expérimentalement retrouvés dans les cinq premiers résidus des profils. Une méthode également basée sur la structure a été développée par Bhattacherjee & Wallin [2013]. L'algorithme Monte Carlo utilisé explore cette fois-ci à la fois l'espace des séquences et des conformations du peptide. La fonction d'énergie utilisée comporte cinq termes énergétiques qui modélisent les interactions du squelette, des chaines latérales, les liaisons hydrogènes et l'effet de désolvatation du squelette. L'espace des séquences des positions P_0, P_{-1} et P_{-2} a été exploré pour trois domaines PDZ : un domaine de classe I (PDZ3 PSD-95), un domaine de classe II (PDZ6 GRIP1) et un domaine liant des peptides des classes I et II (PICK1). Les profils obtenus pour PSD-95 et GRIP1 sont en accord avec leur classe respective, excepté pour la position P_{-2} de PSD-95. Le profil obtenu pour PICK1 a permis d'identifier ce domaine comme appartenant plutôt aux domaines PDZ de classe II.

Les méthodes présentées précédemment prédisent les motifs reconnus par un domaine PDZ. Il peut également être intéressant de prédire l'affinité des complexes PDZ:peptide. Dans cette optique, Kaufmann *et al.* [2011] ont modifié la fonction d'énergie de Rosetta afin de prédire l'affinité des complexes PDZ:peptide. La fonction d'énergie de Rosetta est constituée de six termes énergétiques dont un terme de van der Waals, un terme de solvatation, un terme décrivant la probabilité d'observer la conformation d'une chaine latérale données, un potentiel de paire empirique et un terme de liaison hydrogène. Les poids associés aux différents termes ont été paramétrés puis le modèle a été appliqué à un jeu de 28 peptides liant le domaine PDZ3 de PSD-95 avec des énergies libres comprises entre -8,7 et -4,9 kcal/mol. La modèle obtenu présente une erreur standard de 0,79 kcal/mol et un coefficient de corrélation de 0,66.

1.1.3.3 Dessin computationel de protéine

La modification d'un domaine PDZ ou la création d'un peptide pouvant modifier son activité représente en enjeu thérapeutique important. Plusieurs études portent sur la modification

de domaines PDZ pour changer leur spécificité. Parmi celles-ci, Reina *et al.* [2002] ont modifié le domaine PDZ3 de PSD-95 (classe I) pour qu'il lie des peptides de classe II ou qu'il reconnaisse spécifiquement les positions P_{-1} et P_{-3} d'un peptide. Pour cela, les positions de l'interface (entre 7 et 12 résidus selon les cas) ont été modifiées grâce au programme Perla (López *et al.* [2001]) en présence de trois peptides différents. Ce programme identifie les séquences les plus stables en se basant sur une fonction d'énergie composée de termes physiques et statistiques. Les trois mutants créés lient leur peptide respectif avec une affinité comparable au sauvage dans deux cas et 100 fois supérieure dans le troisième. Les auteurs ont ensuite mis en évidence que dans l'un des trois cas, les mutations introduites correspondent à la séquence naturelle d'un autre domaine PDZ pouvant également lier le peptide. Melero *et al.* [2014] se sont intéressés à la transférabilité des mutations entre domaines PDZ proches. Ils ont dans un premier temps identifié une mutation entrainant un changement de spécificité du résidu reconnu à la position P_{-2} du peptide. Le transfert de la mutation dans six autres domaines n'a permis de changer la spécificité que dans la moitié des cas. Une étude plus poussée a montré que la transférabilité n'était pas possible en raison de subtiles différences structurales.

D'autre études portent sur le développement de peptides inhibiteurs, c'est-à-dire des peptides se liant mieux que le partenaire naturel. Par une approche rationnelle, Xiao *et al.* [2017] ont produit des inhibiteurs peptidiques du domaine PDZ de PTPN4. À partir de simulations de dynamique moléculaire de neuf complexes PTPN4:peptide, couplées à une analyse MM/PBSA, les auteurs ont pu mettre en évidence l'importance des interactions électrostatiques dans la liaison du peptide. En renforçant ces interactions par le biais de trois mutations simples et une mutation double, des peptides inhibiteurs ont pu être produits. Des mesures expérimentales ont par la suite montré que ces peptides se liaient avec une affinité comprise entre 0,15 et 0,86 μM contre 1μM pour le peptide naturel.

Des approches automatiques ont également été appliquées pour identifier des peptides inhibiteurs. Roberts *et al.* [2012] ont créé des peptides inhibant l'interaction de CAL avec son partenaire CFTR. Pour cela, le programme K* a été utilisé pour explorer l'espace des conformations et des séquences du peptide et estimer la constante d'association K_a de chaque variant. Parmi les 2166 peptides produits les 11 meilleures prédictions ont été testées expérimentalement et présentent des affinités comprises entre 2,3 et 18 μM. Ces affinités sont meilleures que celle du ligand naturel de CAL le plus affin (21 μM) et 170 fois meilleures que l'affinité pour CFTR. Zheng *et al.* [2015] se sont intéressés au problème de la sélectivité dans la créa-

tion d'inhibiteur. Leur étude porte sur les domaines PDZ des protéines N2P2 et M3P6 qui présentent des activités oncogéniques opposées, N2P2 étant activatrice et M3P6 inhibitrice. Afin d'identifier des inhibiteurs de N2P2, les auteurs ont développé une méthode basée sur la structure du complexe et prenant en compte sa flexibilité. Afin d'assurer la sélectivité pour le domaine de N2P2, chaque peptide est testé sur N2P2 et sur M3P6. Les meilleurs candidats correspondent aux séquences ayant un bon score pour N2P2 et un mauvais score pour M3P6. Cette méthode a permis d'identifier trois peptides présentant une affinité de l'ordre du micromolaire pour N2P2 et incapables de se lier à M3P6.

1.2 La protéine Tiam1

Cette étude porte principalement sur le domaine PDZ de Tiam1. Pour des raisons de simplicité, lorsque cela n'est pas précisé, Tiam1 fait référence au domaine PDZ et non à la protéine complète. De même, nous ferons référence aux différents peptides par le nom de la protéine dont ils sont issus.

1.2.1 Structure et rôle biologique

La protéine Tiam1 (*T-lymphoma Invasion And Metastasis 1*) a été initialement décrite comme étant capable de conférer un phénotype invasif aux lymphocytes lorsqu'elle était active, ce qui lui a valu son nom (Habets *et al.* [1994]). Cette protéine, longue d'environ 1600 acides aminés est une protéine multi-domaines dont la principale fonction est d'activer les protéines Rac1, Cdc42 et, dans une moindre mesure, RhoA qui appartiennent à la famille des Rho GTPases. Ainsi, en réponse à un stimulus, la protéine Tiam1 va se lier à la GTPase provoquant le détachement du GDP et favorisant la liaison du GTP. Cela a pour effet d'activer la GTPase qui activera à son tour d'autres protéines.

De par son interaction avec Rac1, la protéine Tiam1 est impliquée dans un grand nombre de processus cellulaires tels que la régulation de l'activité du cytosquelette, la migration et l'adhésion cellulaire ou encore la croissance et la survie cellulaire (Boissier & Huynh-Do [2014]). Le rôle de Tiam1 dans ces processus suppose l'existence de mécanismes qui régulent finement son activité. Cette régulation est notamment assurée par la présence de cinq domaines de liaison impliqués dans la reconnaissance des partenaires protéiques à travers des interactions protéine-protéine (figure 1.4). La protéine Tiam1 comprend ainsi un domaine PH impliqué

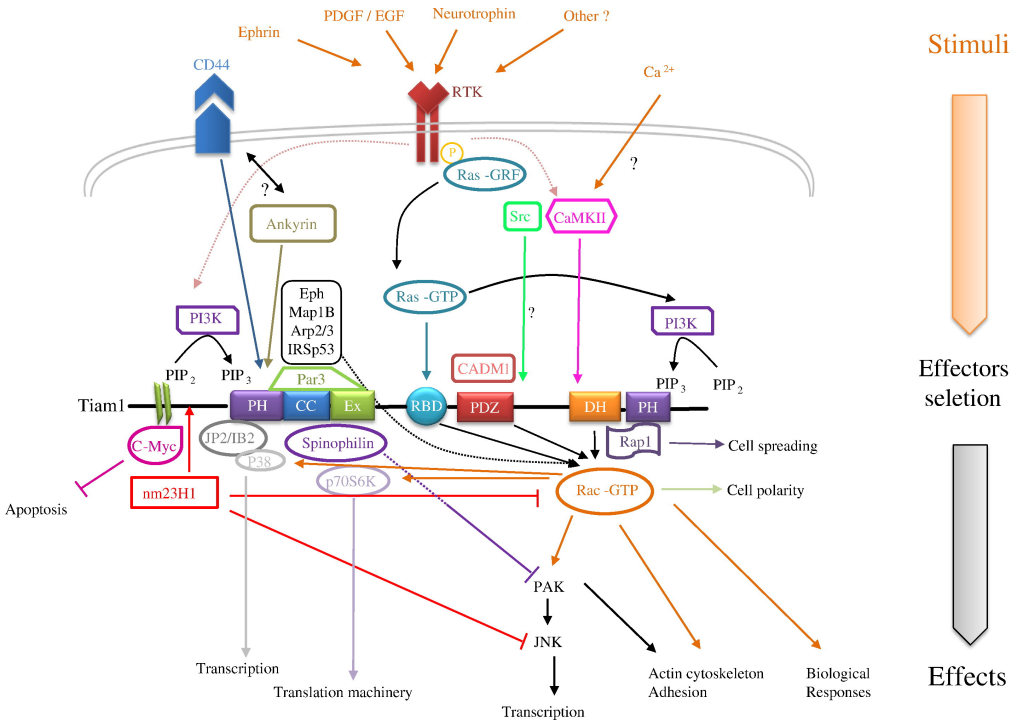


Figure 1.4 – Représentation schématique des precessus cellulaires dépendants de la protéine Tiam1. D'après Boissier & Huynh-Do [2014].

dans son association à la membrane plasmique et dans les interactions avec les protéines du cytosquelette (Spinophiline, JIP2/IB2, Par3), un domaine qui lie les protéines de la famille Ras (RDB), un domaine PDZ reconnaissant plusieurs partenaires (CADM1, Sdc1, Sdc3, Caspr4) qui peut dans certains cas être régulé par la phosphorylation de ses partenaires (Sulka *et al.* [2009] ; Rousselle & Beck [2013]) et la combinaison des domaines PH-DH permettant la liaison des Rho GTPases.

Tiam1 étant un intermédiaire important dans un grand nombre de processsus cellulaires, sa dérégulation, et notamment sa suractivation, est impliquée dans plusieurs types de cancers. En effet, il a été montré que la surexpression de Tiam1, par l'activation de la protéine Rac1, augmentait la motilité des cellules (Hall [1998]). Cet effet pourrait en partie expliquer pourquoi une activité de Tiam1 trop importante favorise le développement de cancers colorectaux (Minard *et al.* [2005, 2006]), des poumons (Adam *et al.* [2001]) et plus généralement la formation de métastases (Minard *et al.* [2004]). Tiam1 constitue donc une excellente cible thérapeutique puisque son inhibition pourrait potentiellement stopper les cascades de signalisation responsables du développement de cellules cancéreuses. Cette inhibition représente cependant un véritable défi au vu du nombre important de processus dans lesquels cette protéine est impliquée.

Tableau 1.1 – Comparaison de séquences reconnues par les domaines PDZ de Tiam1 et Tiam2. D'après Shepherd *et al.* [2011]

Référence	P_{-4}	P_{-3}	P_{-2}	P_{-1}	P_0
Tiam1					
Songyang *et al.* [1997]	[X]	[I]	[FY]	[YH]	[AF]
Tonikian *et al.* [2008]	[F]	[ILM]	[G]	[W]	[F]
Shepherd *et al.* [2011]	[RK]	[IR]	[FY]	[YR]	[ACF]
Tiam2					
Tonikian *et al.* [2008]	[R]	[STE]	[ST]	[SR]	[V]
Shepherd *et al.* [2011]	[K]	[RKH]	[YRKH]	[YH]	[FY]

1.2.2 Partenaires du domaine PDZ de Tiam1

Parmi l'ensemble des partenaires interagissant avec Tiam1, un certain nombre le font par l'intermédiaire du domaine PDZ de Tiam1. Parmi les partenaires connus, on peut citer les protéines de la famille des syndecans (Sdc) et la protéine CADM1. Ces protéines n'étant pas impliquées dans les mêmes cascades de signalisation, leur activation par Tiam1 n'a pas les mêmes effets. La liaison de CADM1 active la protéine Rac1 par le biais de Tiam1 et provoque le réarrangement du cytosquelette ce qui favorise l'infiltration des leucocytes dans les tissus (Masuda *et al.* [2010]). Les protéines de la famille des syndecans sont des récepteurs transmembranaires jouant un rôle dans l'adhésion des cellules. On compte différentes formes de syndecans (numérotées de 1 à 4) dont deux sont reconnues par le domaine PDZ de Tiam1 (1 et 3) (Shepherd *et al.* [2010] ; Liu *et al.* [2013] ; Cheng *et al.* [2016]). Enfin, Caspr4 est une protéine de la famille des Neurexines impliquée dans l'adhésion des cellules (Spiegel *et al.* [2002]). Bien qu'aucune interaction *in vivo* n'ait encore été identifiée entre Caspr4 et Tiam1, des expériences *in vitro* ont prouvé que Tiam1 peut lier l'extrémité C-terminale de Caspr4 (Shepherd *et al.* [2010]). En plus de ces partenaires avérés, huit protéines répertoriées dans la base de données PROSITE (de Castro *et al.* [2006]) ont été identifiées comme partenaires potentiels de Tiam1 d'après la séquence de leur extrémité C-terminale (Shepherd *et al.* [2011]).

1.2.3 Positions impliquées dans la spécificité du domaine PDZ de Tiam1

Le domaine PDZ de Tiam1 est un domaine de classe II et reconnait donc spécifiquement le motif X-Φ-X-Φ. Malgré cette classification, la nature exacte des séquences reconnues reste

incertaine et varie en fonction des études (tableau 1.1). L'analyse des structures de complexes Tiam1:peptide, de données RMN, d'une bibliothèque combinatoire de peptides et de mutants a permis d'identifier les positions responsables de la spécificité de Tiam1 (Shepherd *et al.* [2010, 2011] ; Liu *et al.* [2013, 2016]).

La position P_0 est reconnue par la poche S_0 formée par les résidus Y858, F860, L915 et L920 qui lie les types Ala et Phe. Le groupement carboxylate de l'extrémité C-terminale interagit avec la boucle conservée β_1-β_2 par l'intermédiaire du squelette de la glycine 858 et la lysine 850 (figure 1.5). La position P_{-1} est exposée au solvant et reconnait les types Tyr et Arg au travers d'interactions hydrophobes ou de liaisons hydrogène avec les résidus S861, N876 et S877. Cette position peut également correspondre à une phosphotyrosine, sans modification de l'affinité. Elle est alors reconnue par les résidus T857 et R879 (Liu *et al.* [2013]). La position P_{-2} reconnait spécifiquement les acides aminés aromatiques (Phe et Tyr) au sein de la poche hydrophobe S_{-2} formée par les résidus L911 et K912. La position P_{-3} est exposée au solvant et est donc peu contrainte. Cependant, les types Ile, Leu, Tyr et Arg sont privilégiés par la formation d'interactions hydrophobes avec les résidus à la surface du domaine PDZ et la chaine latérale du résidu P_{-1}. Enfin, la position exposée P_{-4} reconnait les résidus chargés Lys, Arg et Glu. Lorsqu'un Glu est présent, il est stabilisé par des interactions avec S908 et K912. La spécificité des positions P_{-5} à P_{-7} est mal déterminée. La présence d'un résidu chargé à la position P_{-6} semble cependant favorable car il permet la formation d'un réseau de liaisons hydrogène entre les résidus P_{-6}, P_{-3}, P_{-1} et N876 (Liu *et al.* [2013]).

En comparant la séquence de Tiam1 avec celle de son homologue de classe I, Tiam2 (tableau 1.1), Shepherd *et al.* [2011] ont montré que les résidus L915 et L920 de S_0 et L911 et K912 de S_{-2} étaient responsables de la différence de spécificité entre ces deux protéines. Ainsi, la mutation de ces quatre positions vers les types présents chez Tiam2 (L911M, K912E, L915F et L920V) suffit à inverser la spécificité de Tiam1. Ce changement s'explique principalement par la modification du volume de la poche S_0 et de la charge du résidu 912 qui interagit avec les positions P_{-2} et P_{-4}.

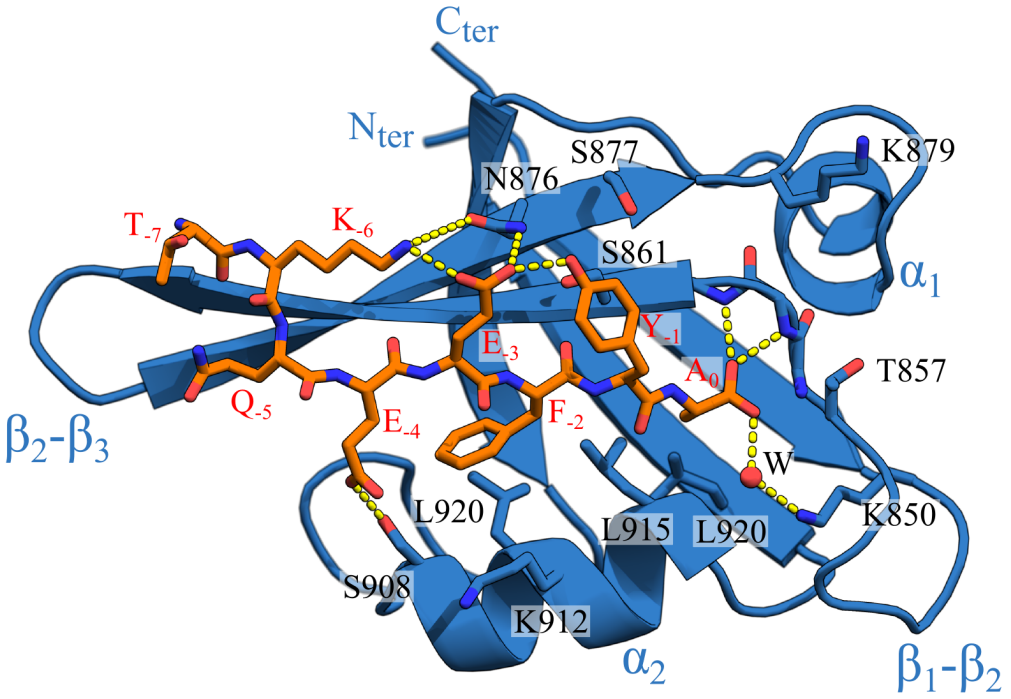


Figure 1.5 – Interface de liaison de Tiam1 avec le peptide Sdc1. Tiam1 est en bleu et Sdc1 en orange. Les résidus impliqués dans la reconnaissance du peptide sont représentés en bâtons. W est une molécule d'eau intervenant dans la liaison de l'extrémité C-terminale du peptide. Les liaisons hydrogènes et les ponts salins sont représentés part les pointillés jaunes

1.2.4 Données disponibles concernant le domaine PDZ de Tiam1

1.2.4.1 Données structurales disponibles

Huit structures du domaine PDZ de Tiam1 sont disponibles dans la PDB (tableau 1.2), sept structures cristallographiques et une structure RMN. Toutes les structures cristallographiques présentent une résolution comprise entre 1,3 et 2,3 Å. Parmi ces structures, on retrouve le domaine PDZ de Tiam1 sous sa forme apo, c'est-à-dire sans ligand, et en complexe avec différents peptides. La structure d'un quadruple (QM) mutant a également été résolue. Il correspond aux quatres mutations de spécificité présentées précédemment (L911M, K912E, L915F et L920V).

1.2.4.2 Affinités expérimentales

L'affinité du domaine PDZ de Tiam1 a été mesurée pour 50 complexes par anisotropie de fluorescence (Shepherd *et al.* [2010, 2011] ; Liu *et al.* [2013, 2016]). Ces mesures ont été

Tableau 1.2 – Structures cristallographiques et NMR du domaine PDZ de Tiam1 disponibles dans la PDB.

PDB ID	Méthode	Séquence PDZ	Peptide	Séquence Peptide	Résolution (Å)	Référence
3KZD	X-ray	WT	apo	-	1,3	Shepherd *et al.* [2010]
3KZE	X-ray	WT	consensus	SSRKEYYA	1,8	Shepherd *et al.* [2010]
4GVC	X-ray	WT	pSdc1	TKQEEFpYA	1,5	Liu *et al.* [2013]
4GVD	X-ray	WT	Sdc1	TKQEEFYA	1,9	Liu *et al.* [2013]
4NXP	X-ray	QM	apo	-	2,3	Liu *et al.* [2016]
4NXQ	X-ray	QM	Caspr4	ENQKEYFF	2,1	Liu *et al.* [2016]
4NNR	X-ray	QM	Neurexine	NKEKDYYV	1,8	Liu *et al.* [2016]
2D8I	RMN	WT	apo	-	-	Qin *et al.* [2005]

Figure 1.6 – Bibliothèque combinatoire des types reconnus par Tiam1 au niveau des cinq positions C-terminales du peptide. Chaque acide aminé est représenté par son code à une lettre dont la taille est proportionnelle à son occurrence. Les données sont issues de l'étude effectuée par Shepherd *et al.* [2011].

effectuées sur le domaine PDZ isolé avec des peptides de huit acides aminés. Parmi les systèmes étudiés, 25 correspondent à la forme sauvage de Tiam1 liée à différents peptides. L'autre moitié regroupe des mutants du domaine PDZ de Tiam1 au niveau des quatre positions impliquées dans la spécificité de Tiam1 et Tiam2 (Shepherd *et al.* [2011]) et d'une lysine favorisant la liaison de la forme phosphorylée de Sdc1 (Liu *et al.* [2013]). Deux complexes ont des valeurs de K_d très élevées ($\geq$ 1600 μM) et six ont des affinités non mesurables et sont de ce fait considérés comme non liants (tableau 1.3). Tous les autres peptides se lient à Tiam1 avec une valeur de K_d comprise entre 10,8 et 453 μM ce qui représente une plage d'énergie libre (ΔG) de 2,2 kcal/mol.

En complément des affinités mesurées, une bibliothèque combinatoire de peptides reconnus par Tiam1 a été déterminée expérimentalement (Shepherd *et al.* [2011]). Dans cette étude, environ 290000 variants peptidiques au niveau des positions P_0 à P_{-4} ont été testés. Cela a permis d'identifier 70 séquences liant Tiam1. À partir de ces peptides un profil des types préférentiellement reconnus aux différentes positions a été établi (figure 1.6).

Tableau 1.3 – Affinités expérimentales des complexes Tiam1:peptide.

complexe	séquence	K_d (μM)		ΔG_b (kcal/mol)	
Sdc1	TKQEEFYA	26,9	± 0,9	-6,23	± 0,02
pSdc1	TKQEEFpYA	19,3	± 1,5	-6,44	± 0,05
Sdc1.A0F	TKQEEFYF	55,7	± 3,6	-5,81	± 0,04
Sdc1.A0M	TKQEEFYM	378	-	-4,67	-
Sdc1.F2I	TKQEEIYA	105	-	-5,43	-
Sdc1.F2N[a]	TKQEENYA	>400	-	>-4.64	-
Sdc1.E4K	TKQKEFYA	106	± 7	-5,42	± 0,04
Sdc1.E4L	TKQLEFYA	69,5	-	-5,67	-
Sdc1.E3D,Y1T	TKQEDFTA	118	-	-5,36	-
Sdc1.E3T,Y1K	TKQETFKA	253	-	-4,91	-
Sdc2	APTKEFYA	453	± 22	-4,56	± 0,03
Sdc3	DKQEEFYA	33,4	± 1,9	-6,11	± 0,03
Sdc4	APTNRFYA	397	± 17	-4,64	± 0,03
cons	SSRKEYYA	112	± 5	-5,39	± 0,03
Caspr4	ENQKEYFF	19	± 0,4	-6,44	± 0,01
Caspr4.F0A	ENQKEYFA	64,8	± 5,9	-5,72	± 0,05
Neu	NKDKEYYV	2400	± 250	-3,58	± 0,07
CADM1	EEKKEYFI	1600	± 100	-3,82	± 0,04
YAAEKYWA	YAAEKYWA	90,9	± 8,9	-5,52	± 0,06
YAAKAFRF	YAAKAFRF	200	± 50	-5,07	± 0,16
YAAYRYRA[a]	YAAYRYRA	>250	-	>-4,91	-
YAARKFAK[a]	YAARKFAK	>250	-	>-4,91	-
YAAGRKHF[a]	YAAGRKHF	>250	-	>-4,91	-
YAALIHKF[a]	YAALIHKF	>250	-	>-4,91	-
YAAQKHFH[a]	YAAQKHFH	>250	-	>-4,91	-
Variants de la protéine					
K879:Sdc1		64,7	± 0,1	-5,72	± 0,00
K879:pSdc1		170	± 6	-5,14	± 0,02
L911M:Sdc1		34,7	± 0,7	-6,09	± 0,01
K912E:Sdc1		140	± 20	-5,27	± 0,09
L911M,K912E:Sdc1		211	± 36	-5,02	± 0,10
L915F:Sdc1		81	± 7	-5,58	± 0,05
L920V:Sdc1		46	± 3	-5,92	± 0,04
L915F,L920V:Sdc1		250	± 20	-4,92	± 0,05
QM:Sdc1		122	± 8	-5,34	± 0,04
QM:Sdc1.A0F		39	± 2	-6,02	± 0,03
L911M:Caspr4		14	± 0,3	-6,62	± 0,01
K912E:Caspr4		58,6	± 4,1	-5,78	± 0,05
L911M,K912E:Caspr4		28,9	± 0,9	-6,19	± 0,02
L915F:Caspr4		61	± 4	-5,75	± 0,04
L920V:Caspr4		10,8	± 0,6	-6,78	± 0,03
L915F,L920V:Caspr4		76,5	± 3,4	-5,62	± 0,03
QM:Caspr4		18,3	± 0,3	-6,46	± 0,01
QM:Caspr4.F0A		170	± 9	-5,14	± 0,03
L911M,K912E:Neu		270	± 150	-4,98	± 0,37
L915F,L920V:Neu		166	± 12	-5,16	± 0,05
QM:Neu		46	± 2	-5,92	± 0,07
QM:CADM1		118	± 9	-5,36	± 0,05
QM:Sdc2		100	± 3	-5.46	± 0,01
QM:Sdc3		138	± 2	-5,27	± 0.00
QM:Sdc4		209	± 17	-5,02	± 0,05

[a] : Affinité au delà des limites de détection de la méthode.

Deuxième partie

Étude du domaine PDZ de Tiam1 par des approches de dessin computationel de protéine

Chapitre 2

Le dessin computationel de protéine

Le dessin de protéine permet, par des approches expérimentales et computationnelles, de modifier les propriétés structurales ou fonctionnelles d'une protéine en modifiant sa séquence en acides aminés. Il peut par exemple être utilisé pour augmenter la stabilité d'une protéine ou encore modifier son affinité pour un ligand. La structure des protéines étant directement liée à leur séquence, les modifications apportées doivent être choisies de sorte à ne pas altérer le repliement de la protéine cible. Le dessin de protéine cherche donc à identifier les séquences présentant les propriétés recherchées et compatibles avec la structure de la protéine d'intérêt.

Deux stratégies sont couramment adoptées expérimentalement. La première, dite rationnelle, se base sur la connaissance de la structure tridimensionnelle de la protéine pour identifier les régions dont la modification pourrait apporter les propriétés recherchées. Cette approche s'appuie sur les techniques de mutagenèse dirigée, aujourd'hui largement utilisées. La deuxième est dite combinatoire et consiste à introduire aléatoirement des mutations dans la protéine cible puis à appliquer une procédure de sélection afin d'identifier les mutants possédant les propriétés recherchées.

Le problème peut également être traité par ordinateur ce qui permet d'étudier de façon rapide, exhaustive et peu couteuse les mutants d'une protéine. Cette approche est appelée dessin computationnel de protéine ou CPD (pour *Computational Protein Design*). Elle nécessite de connaitre la structure de la protéine d'intérêt. De nombreux programmes de CPD ont été développés ces dernières décennies. Ils sont basés sur trois ingrédients communs : la définition de l'espace conformationnel, un algorithme d'exploration pour échantillonner les séquences et une fonction d'énergie capable de discriminer les séquences générées.

Dans ce chapitre nous nous intéressons aux différentes composantes des programmes de CPD. Nous détaillerons ensuite plus en détails le fonctionnement des deux programmes qui seront utilisés par la suite, Proteus et Rosetta.

2.1 Modélisation de l'espace conformationnel

Les protéines sont des objets flexibles, constamment en mouvement et explorant un espace continu de conformations. Cette flexibilité jouant un rôle important dans les processus biologiques et leur régulation, elle devrait être prise en compte lors de l'exploration des séquences. Malheureusement, cet espace conformationnel est trop vaste pour pouvoir être modélisé dans son intégralité. et doit donc être simplifié. Cette simplification peut se faire en discrétisant l'espace des conformations du squelette de la protéine mais également en discrétisant celui des chaines latérales. De plus, afin d'évaluer la stabilité des séquences générées, les programmes de CPD estiment la différence d'énergie libre entre les états replié et déplié de la protéine. L'état déplié n'ayant pas de structure définie, il faut développer un modèle simplifié permettant d'évaluer son énergie. Dans cette partie, nous nous intéressons aux différentes méthodes utilisées pour décrire l'espace conformationnel des protéines dans leurs états replié et déplié.

2.1.1 Modélisation de l'état replié

2.1.1.1 Modélisation des chaines latérales

Dans l'état replié, les chaines latérales des acides aminés adoptent préférentiellement un petit ensemble de conformations (Finkelstein & Ptitsyn [1977] ; Janin *et al.* [1978] ; Ponder & Richards [1987]). Ces conformations, énergétiquement favorables, sont appelées rotamères et permettent de représenter l'espace des conformations de chaque type d'acide aminé de manière discrète. Chaque rotamère est ainsi décrit par des valeurs d'angles de torsions particulières (figure 2.1). Les rotamères sont regroupés au sein de bibliothèques construites à partir d'études statistiques des structures de protéines connues. Ces bibliothèques ont été développées selon différents critères. Elles peuvent être indépendantes du squelette de la protéine (Tuffery *et al.* [1991] ; De Maeyer *et al.* [1997]), dépendantes des angles ϕ et ψ (Dunbrack & Karplus [1993] ; Dunbrack & Cohen [1997] ; Towse *et al.* [2016]) ou dépendantes des structures secondaires α ou β (Schrauber *et al.* [1993] ; Lovell *et al.* [2000]). Certains modèles utilisent également des rotamères continus. Dans ce cas, chaque rotamère peut explorer une partie de l'espace des angles χ (Gainza *et al.* [2012]).

Figure 2.1 – Angles χ d'une arginine (A) et exemples de rotamères associés (B).

2.1.1.2 Modélisation du squelette

Dans la plupart des programmes de CPD le squelette de la protéine est fixe. Les atomes C, N, C_α et O du squelette ainsi que l'atome C_β sont maintenus rigides. Cette méthode présente quelques limites. En effet, il arrive que certaines chaines latérales soient énergétiquement défavorables alors qu'un léger ajustement du squelette suffirait à diminuer considérablement leur énergie. Ce type de limitation a notamment été démontré par les travaux sur le dessin du cœur du lysozyme T4 par Hurley *et al.* [1992], Baldwin *et al.* [1993] et Mooers *et al.* [2003].

Introduire de la flexibilité au niveau du squelette augmente considérablement l'espace conformationnel. Néanmoins cette flexibilité permet de décrire de manière plus juste la protéine et peut donc améliorer la qualité des séquences produites. Nous présentons ici quelques-unes des approches utilisées pour augmenter l'espace conformationnel du squelette.

Desjarlais & Handel [1995] ont modélisé la flexibilité du squelette à l'aide d'une méthode d'exploration de séquences Monte Carlo couplée à un algorithme génétique permettant d'optimiser la structure de la protéine. Cette approche a permis de modifier jusqu'à huit positions du cœur hydrophobe de la protéine 434 cro. Une seconde étude a montré que l'ajout de flexibilité dans le squelette n'améliorait pas les résultats de prédictions de stabilité (Desjarlais & Handel [1999]). Ce résultat peut s'expliquer par le fait que le cœur des protéines est une région géométriquement très contrainte, avec peu de degrés de liberté.

Su & Mayo [1997], en se basant sur les travaux de Harbury *et al.* [1995] et Offer & Sessions [1995], abordèrent le problème en traitant les structures secondaires comme des corps rigides capables d'adopter des positions différentes les uns par rapport aux autres. Jusqu'à six positions du cœur de la protéine Gβ1 ont été redessinées avec cette méthode.

Harbury *et al.* [1998] se sont intéressés aux protéines présentant une symétrie telles que les *coiled-coils* ou les *TIM barrels*. La symétrie permet en effet de décrire leur squelette par des

équations paramétriques. Cette méthode fût utilisée pour modéliser une famille non naturelle de structures de type *right-handed coiled-coil*. Elle n'est cependant pas applicable à la majorité des protéines, puisque la plupart ne possèdent pas de symétrie.

Le groupe de Baker proposa une méthode consistant à échantillonner alternativement l'espace des séquences et des structures. Les séquences sont d'abord optimisées pour un squelette fixe, puis la conformation du squelette est optimisée à séquence fixe (Saunders & Baker [2005]). Grâce à cette méthode,Kuhlman *et al.* [2003] ont créé un nouveau pli protéique, validé par la suite expérimentalement.

Une autre approche développée par Smith & Kortemme [2008] consiste à appliquer un mouvement dit de *backrub* au squelette de la protéine. Ce mouvement correspond au déplacement de l'atome C_{α_i} d'un résidu i ainsi que sa chaine latérale autour d'un axe formé par les atomes $C_{\alpha_{i-1}}$ et $C_{\alpha_{i+1}}$. Cette approche a été utilisée par Georgiev *et al.* [2008] pour redessiner deux protéines (GrsA-PheA et Gβ1). Les auteurs ont montré que l'utilisation du mouvement *backrub* permet d'explorer des séquences de plus basses énergies.

Récemment, Druart *et al.* [2016] ont développé une méthode qui prend en compte la flexibilité de certaines régions de la protéine. Pour cela, une petite bibliothèque de squelettes est d'abord produite par dynamique moléculaire. Une simulation Monte Carlo hybride explore ensuite à la fois l'espace des conformations des chaines latérales et du squelette. Cette approche a été appliquée au design d'une boucle flexible située dans le site actif de la tyrosyl-ARNt synthétase.

2.1.2 Modélisation de l'état déplié

La stabilité d'une protéine dépend de la différence d'énergie libre entre son état déplié et son état replié. Afin de calculer la stabilité des séquences produites par CPD il est donc nécessaire de modéliser l'état déplié de la protéine. Cet état est extrêmement difficile à modéliser puisqu'il est peu structuré et correspond à une distribution continue de conformations ou micro-états d'énergies similaires.

En supposant que l'énergie de repliement ne dépend que de la séquence en acides aminés, on peut cependant décrire cet état de manière implicite. Une possibilité est de maintenir fixe la composition en acides aminés ou bien la proportion hydrophile-hydrophobe. De cette manière l'état déplié devient équivalent pour toutes les séquences et n'entre plus en jeu lors de la

comparaison des différentes séquences. Ces approches ont été appliquées par Dahiyat & Mayo [1997] et Koehl & Levitt [1999] mais elles ne sont pas optimales car elles contraignent l'espace des séquences exploré.

La méthode la plus utilisée dans les programmes de CPD consiste à définir une énergie de référence pour chaque type d'acide aminé, notée E_X. Ce modèle fait ainsi l'hypothèse que, dans l'état déplié, les chaines latérales des résidus n'interagissent pas entre elles. On peut alors calculer chaque énergie de référence en considérant un acide aminé entièrement exposé au solvant. Cette énergie est généralement calculée dans un tripeptide où le résidu X est encadré par deux alanines (Dahiyat & Mayo [1996] ; Wernisch *et al.* [2000]). L'énergie totale de l'état déplié est ensuite calculée en sommant les énergies de référence selon les types présents dans la séquence. Les énergies de référence influent directement sur la composition des séquences explorées et doivent donc être choisies avec soin. Si par exemple les énergies de référence ne sont pas prises en compte, cela conduit à une composition fantaisiste (Suárez & Jaramillo [2009]).

2.2 Fonction d'énergie

Pour comparer les conformations générées par le programme de CPD, il faut définir une fonction d'énergie. Elle doit être suffisamment juste pour capturer les interactions interatomiques de la protéine tout en étant rapide à calculer. Cette fonction est généralement basée sur les fonctions de la dynamique moléculaire auxquelles s'ajoutent parfois des termes issus d'analyses statistiques. Dans le cas du CPD, lorsque l'espace conformationnel est discrétisé, la fonction d'énergie est généralement décomposable par paire de résidus. Comme nous le verrons un peu plus loin, cette propriété permet de déterminer rapidement l'énergie du système en sommant les énergies de chaque paire.

2.2.1 Fonction d'énergie issue de la mécanique moléculaire

En mécanique moléculaire, le système est décrit comme un ensemble de particules sphériques reliées par des ressorts. L'énergie d'une protéine est composée de deux termes : un terme décrivant les interactions des atomes de la protéine entre eux (E_{MM}) et un terme décrivant

l'effet du solvant sur la protéine (E_{solv}) :

$$E = E_{MM} + E_{solv} \tag{2.1}$$

Le terme E_{MM} peut lui-même se décomposer en deux parties, un terme lié qui correspond aux interactions des atomes séparés par une à trois liaisons covalentes, et un terme non-lié qui correspond aux autres paires d'atomes.

2.2.1.1 Énergie d'interactions liées

L'énergie d'interactions liées correspond aux interactions entre les atomes distants de moins de trois liaisons covalentes :

$$E_{liées} = E_{liaison} + E_{angle} + E_{dièdre} + E_{impropre} \tag{2.2}$$

$E_{liaison}$ correspond à l'élongation des liaisons, E_{angle} à la déformation des angles, $E_{dièdre}$ à la torsion des angles dièdres et $E_{impropre}$ à la déformation de groupes plans. (figure 2.2)

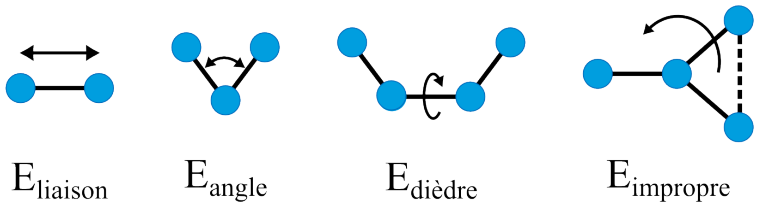


Figure 2.2 – Les quatre termes contribuant aux interactions liées. Les cercles et les lignes représentent respectivement les atomes et les liaisons covalentes.

2.2.1.2 Énergie d'interactions non liées

En mécanique moléculaire, les interactions non-liées sont prises en compte lorsque les atomes sont distants de plus de trois liaisons covalentes. Ces interactions sont généralement séparées en deux composantes, les interactions de type van der Waals (E_{vdW}) et les interactions électrostatiques (E_{elec}) (figure 2.3).

Énergie de van der Waals Les interactions de van der Waals sont des interactions électrostatiques de faible intensité entre deux atomes. Ces interactions sont la plupart du temps

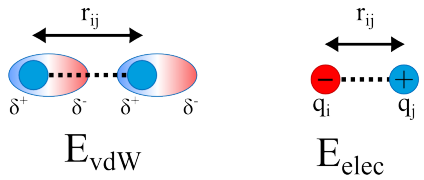


Figure 2.3 – Représentation schématique des deux termes contribuant aux interactions non liées. Les cercles représentent les atomes.

modélisées par un potentiel de Lennard-Jones :

$$E_{vdW}(i,j) = 4\epsilon[(\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}] \tag{2.3}$$

où ϵ et σ sont des constantes et r_{ij} est la distance entre les atomes i et j. Le premier terme modélise les forces répulsives de Pauli et domine lorsque les atomes sont proches. Le deuxième terme modélise les forces attractives entre dipôles instantanés et domine lorsque les atomes sont éloignés.

Énergie électrostatique L'énergie électrostatique est modélisée par un terme de Coulomb entre charges partielles atomiques. Ce terme dépend de la distance entre les charges et de l'écrantage diélectrique du milieu :

$$E_{elec} = \frac{q_i q_j}{\epsilon r_{ij}} \tag{2.4}$$

où q_i, q_j représentent les charges, ϵ la constante diélectrique du milieu et r_{ij} la distance entre les atomes i et j.

2.2.1.3 Modélisation implicite du solvant

Le solvant, de par sa constante diélectrique élevée, joue un rôle important en écrantant les interactions électrostatiques. Sa modélisation explicite est trop couteuse pour être appliquée au CPD. Le solvant est donc modélisé de manière implicite. L'énergie de solvatation (E_{solv}) comprend deux termes : un terme E_{solv}^{pol} qui décrit les interactions polaires et un terme E_{solv}^{apol} qui décrit l'effet hydrophobe :

$$E_{solv} = E_{solv}^{pol} + E_{solv}^{apol} \tag{2.5}$$

Dans le CPD, la protéine est souvent définie comme un corps de faible constante diélectrique entourée d'un milieu continu ayant une constante diélectrique élevée. La limite entre ces deux

régions est déterminée par la surface moléculaire. Le terme apolaire est généralement modélisé par un terme surfacique :

$$E_{solv}^{surf} = \sum_i \sigma_i A_i \tag{2.6}$$

A_i correspond à la surface accessible au solvant de l'atome i et σ_i à un coefficient d'énergie de surface (en kcal/mol/Å^2) qui dépend de l'hydrophobicité de l'atome (Wesson & Eisenberg [1992]).

Modèle CASA (Coulombic Accessible Surface Area) Le modèle CASA utilise une constante diélectrique ϵ pour pondérer le terme de Coulomb et mimer l'effet d'écrantage du solvant. À ce terme s'ajoute le terme surfacique E_{solv}^{surf} :

$$E_{solv}^{CASA} = (\frac{1}{\epsilon} - 1)E_{elec} - E_{solv}^{surf} \tag{2.7}$$

Ce modèle a donné de bons résultats dans le développement de protéines plus stables et le dessin de cœurs hydrophobes (Dahiyat & Mayo [1996] ; Raha *et al.* [2000]). Cependant, l'utilisation d'une constante diélectrique unique le rend moins adapté au dessin de la surface des protéines.

Modèle de Poisson-Boltzmann Le modèle de Poisson-Boltzmann (PB) est actuellement considéré comme le meilleur modèle de solvant implicite et présente l'avantage d'être fondé sur des concepts physiques. Il décrit la protéine comme une cavité de faible constante diélectrique entourée d'un milieu de forte constante diélectrique. Ce modèle permet de prendre en compte à la fois les fortes interactions électrostatiques entre les groupes chargés et le solvant polarisé, mais également le phénomène d'écrantage du solvant sur les interactions au sein de la protéine. Bien qu'il existe une version décomposable par paires du modèle PB (Marshall *et al.* [2005] ; Vizcarra *et al.* [2008]), cette méthode reste couteuse en temps de calcul. D'autres approches lui sont donc préférées dans le cadre du CPD.

Modèle de Born généralisé Le modèle de Born généralisé (ou GB pour *Generalized Born*) est une approximation du modèle PB (Born [1920]). Dans le modèle GB, les atomes sont modélisés par des sphères ayant une constante diélectrique plus faible que l'environnement. L'effet d'écrantage appliqué à chaque atome i dépend directement de sa distance au solvant.

Cette distance, appelée rayon de Born, reflète l'enfouissement de la charge dans la protéine. L'énergie a la forme :

$$E_{elec}^{GB} = \sum_{ij} \frac{\tau q_i q_j}{2} (r_{ij} + b_i b_j e^{-r_{ij}^2/4b_i b_j})^{-1/2} \tag{2.8}$$

avec $\tau = \frac{1}{\epsilon_{ext}} - \frac{1}{\epsilon_{int}}$, r_{ij} la distance entre les charges q_i et q_j, et b_i le rayon de Born de l'atome i.

2.2.2 Décomposition de l'énergie par paires pour le CPD

Lors de l'exploration des séquences, il est faut évaluer rapidement et efficacement l'énergie. Une approche consiste à précalculer l'ensemble des interactions entre paires de rotamères de manière indépendante (Dahiyat & Mayo [1997] ; Gaillard & Simonson [2014]). L'énergie totale du système est ensuite calculée en sommant l'ensemble des énergies de paires. Cela suppose que l'énergie a la forme :

$$E_{totale} = \sum_{i}^{N} E_{ii} + \sum_{ij}^{N} E_{ij} \tag{2.9}$$

E_{ii} est l'énergie d'interaction du résidu i avec lui-même et avec le squelette de la protéine, E_{ij} l'énergie d'interaction entre les résidus i et j et N le nombre de résidus. L'ensemble des termes énergétiques sont rassemblés dans une matrice carrée. La diagonale contient les termes E_{ii} pour chaque rotamère de chaque position, tandis que les termes E_{ij} sont hors-diagonaux. Cette matrice est ensuite lue au cours de l'exploration des séquences pour déterminer l'énergie des conformations.

La décomposition par paires du terme E_{MM} est exacte, ce qui n'est pas le cas du terme E_{solv} lorsque le GB est utilisé. En effet, l'interaction GB entre deux atomes ne dépend pas uniquement de ces deux atomes mais également des atomes environnants. Pour pallier ce problème, une approximation est réalisée en calculant l'énergie électrostatique de chaque paire dans l'environnement natif (ou NEA pour *Native environnent approximation*). Dans la méthode NEA, les rayons de Born pour un résidu sont ainsi évalués en supposant que les autres atomes sont dans leur conformation native.

Une approximation est également nécessaire pour le terme E_{solv}^{surf}. En effet, la surface enfouie d'une paire de résidus peut être recouverte par la chaine latérale d'un troisième résidu qui n'est pas prise en compte lors du calcul. La surface commune aux trois résidus est alors comptabilisée pour chaque paire, ce qui aboutit à une surestimation de la surface enfouie (figure 2.4). Pour

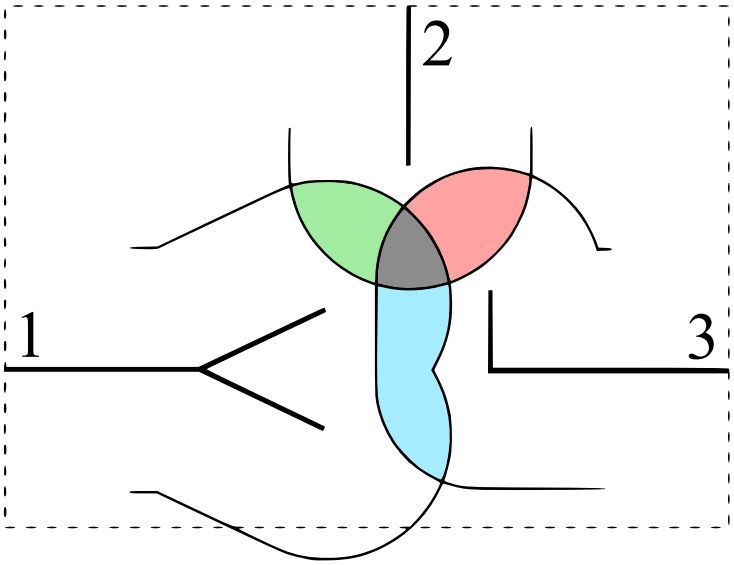


Figure 2.4 – Représentation de trois résidus et de leurs surfaces de contact respectives. Vert : surface de contact entre les résidus 1 et 2 ; rouge : surface de contact entre les résidus 2 et 3 ; bleu : surface de contact entre les résidus 1 et 3 ; gris : surface de contact commune aux trois résidus et comptabilisée plusieurs fois. D'après Street & Mayo [1998]

limiter cette surestimation, un facteur de correction de l'enfouissement peut être appliqué (Street & Mayo [1998]).

2.3 Algorithmes d'explorations

La méthode d'échantillonnage est un point fondamental dans la procédure de CPD. Différents algorithmes d'exploration des séquences et des conformations ont été développés. Ils peuvent être séparés en deux classes : les méthodes stochastiques ou heuristiques et les méthodes déterministes ou exactes. Les méthodes stochastiques ou heuristiques sont basées sur l'aléatoire. Elles ne garantissent pas de trouver la séquence de plus basse énergie et deux simulations indépendantes n'aboutiront pas aux mêmes résultats. Les méthodes déterministes ou exactes permettent quant à elles d'identifier la séquence et la structure de meilleure énergie appelée *Global Minimum Energy Conformation* (GMEC).

2.3.1 Algorithmes stochastiques ou heuristiques

Monte Carlo La méthode de Monte Carlo, initialement introduite par Metropolis *et al.* [1953], échantillonne de manière aléatoire les séquences et les structures. Elle se base sur un

critère d'acceptation ou de rejet d'une conformation. La séquence en acides aminés est d'abord initialisée aléatoirement. Elle est ensuite modifiée soit par un changement de rotamère, soit par une mutation. L'énergie de la nouvelle séquence, E_{new}, est alors comparée à l'énergie de l'état précédent, E_{old}. Si E_{new} est plus faible que E_{old}, la nouvelle conformation est plus favorable que la précédente et est acceptée. Sinon, la modification est acceptée avec une probabilité $p = \exp(\frac{E_{new}-E_{old}}{k_B T})$, où k_B est la constante de Boltzmann et T la température. Grâce à ce critère, il est possible d'accepter des conformations ayant une énergie défavorable, permettant ainsi de passer les barrières énergétiques et de visiter différents minimums locaux. En jouant sur la valeur de la température, on peut augmenter ou diminuer la probabilité d'accepter les conformations d'énergies défavorables.

Algorithme génétique L'algorithme génétique a été proposé par Holland [1975] et repose sur les principes de l'évolution génétique, c'est-à-dire la mutation, la recombinaison et la sélection. Un groupe de séquences, appelé population, est d'abord initialisé aléatoirement. Les séquences sont ensuite triées selon leur énergie. Celles de meilleures énergies sont sélectionnées comme séquences parentes puis modifiées selon les principes cités précédemment pour produire la génération suivante. Un nouveau cycle est ensuite initié en sélectionnant les séquences de plus basses énergies.

Algorithme heuristique de Wernisch L'algorithme de Wernisch *et al.* [2000] identifie la séquence de plus basse énergie ainsi qu'un ensemble de séquences proches. La séquence est d'abord initialisée en assignant aléatoirement un type et un rotamère à chaque position. À chaque itération une position i est choisie aléatoirement et le rotamère optimal pour cette position est sélectionné en tenant compte des chaines latérales environnantes. La procédure est répétée jusqu'à convergence des énergies. Cette méthode ne garantit pas de trouver le minimum global mais un minimum proche de la séquence initiale.

2.3.2 Algorithmes déterministes ou exacts

Dead-End Elimination La méthode du Dead-End Elimination (DEE) élimine au fur et à mesure de l'exploration les rotamères de mauvaises énergies jusqu'à ne conserver qu'un seul rotamère par position (Desmet *et al.* [1992]). Il est possible d'éliminer soit un rotamère, soit une paire de rotamères (Goldstein [1994]). Le premier critère de Goldstein consiste à éliminer

un rotamère r à une position donnée s'il existe, à cette même position, un autre rotamère t tel que sa plus mauvaise contribution énergétique est supérieure à la meilleure contribution énergétique de r. De la même manière, une paire de rotamères i et j peut être éliminée, selon le second critère de Goldstein, s'il existe une autre paire de rotamères r et s aux mêmes positions qui possède une contribution énergétique systématiquement plus faible que la paire i et j. Le principal avantage de cette méthode est la garantie de converger vers le minimum global du système, notamment quand celui-ci est de petite taille.

La théorie du champ moyen L'idée principale de la théorie du champ moyen est de réduire un problème multi-états à un problème mono-état. Pour cela, l'énergie d'interaction d'un rotamère avec tous les autres rotamères de la protéine est résumée à une énergie moyenne (Koehl & Delarue [1994] ; Kono & Doi [1994] ; Lee [1994]). Cette méthode utilise une représentation où chaque rotamère interagit avec toutes les conformations possibles des chaines latérales environnantes, pondérées par leur probabilité respective. De manière itérative, l'énergie moyenne de chaque position est réévaluée jusqu'à convergence des énergies. Une fois la convergence atteinte, les rotamères les plus probables sont positionnés sur la structure du squelette.

2.4 Les principaux programmes de CPD

De nombreux programmes de CPD ont été développés au cours des dernières décennies. Nous pouvons citer ORBIT (Dahiyat & Mayo [1996]), Toulbar2 (Allouche *et al.* [2014]), PocketOptimizer (Masili *et al.* [2012]), Proteus (Schmidt Am Busch *et al.* [2008]), FASTER (Hom & Mayo [2005]), OSPREY (Gainza *et al.* [2013]) et la suite ROSETTA (Kuhlman *et al.* [2003]). Tous ces programmes diffèrent par l'algorithme d'exploration utilisé, la prise en compte de la flexibilité du squelette de la protéine ou encore la fonction d'énergie utilisée. Dans cette partie nous détaillerons uniquement les deux programmes que nous utiliserons, à savoir Proteus et ROSETTA.

2.4.1 Le programme Proteus

Proteus est un programme de CPD développé par l'équipe de Thomas Simonson (Schmidt Am Busch *et al.* [2008] ; Simonson *et al.* [2013] ; Polydorides *et al.* [2016]). Le programme se compose de trois éléments principaux (figure 2.5) :

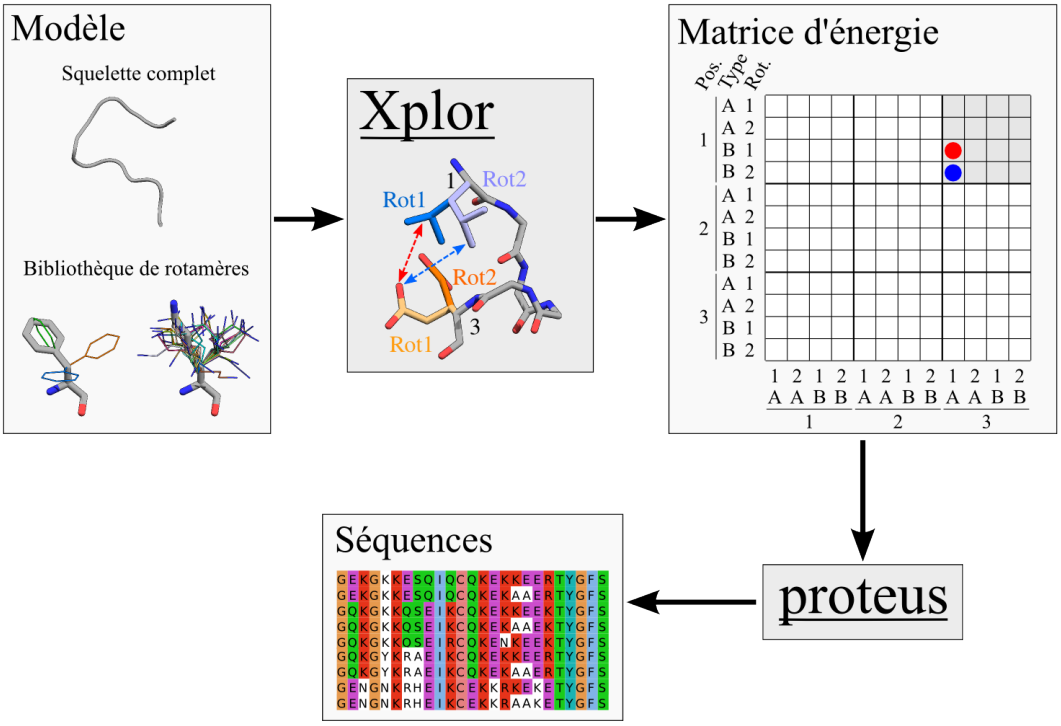


Figure 2.5 – Architecture du programme Proteus. À partir du modèle protéique décomposé en deux parties (squelette et bibliothèque de rotamères), les énergies d'interaction de paires de rotamères sont calculées avec le programme Xplor puis stockées dans la matrice d'énergie. Cette matrice est ensuite lue par la programme proteus afin d'explorer l'espace des séquences.

— un programme de mécanique moléculaire, Xplor (Brünger [1992]), qui calcule les énergies d'interaction au sein du système.

— un ensemble de scripts écrits dans le langage de script de Xplor qui calculent la matrice d'énergie du système.

— un programme C, appelé proteus, qui explore l'espace des séquences et des conformations.

Proteus utilise la bibliothèque de rotamères indépendante du squelette développée par Tuffery *et al.* [1997].

2.4.1.1 Fonction d'énergie

La fonction d'énergie de Proteus est exclusivement basée sur des termes physiques issus de la mécanique moléculaire. Les énergies d'interactions interatomiques sont donc décrites pas un terme d'énergie liée (pour l'énergie interne de chaque rotamère) et un terme pour les interactions non liées. Initialement, l'énergie de solvatation était modélisée par un terme CASA mais elle est maintenant remplacée par un terme GBSA.

2.4.1.2 Algorithmes d'exploration

Différents algorithmes d'exploration sont proposés par le programme proteus : Monte Carlo, algorithme de Wernisch ou champ moyen. Récemment, une méthode Monte Carlo avec échange de répliques a été implémentée dans proteus (Mignon & Simonson [2016]). Elle effectue en parallèle plusieurs simulations Monte Carlo, à des températures différentes, qui échangent à intervalles réguliers leurs conformations. Les simulations à hautes températures pouvant passer des barrières énergétiques plus importantes, elles sont susceptibles d'explorer une plus grande partie du paysage énergétique. Les simulations à plus basses températures explorent quant à elles les conformations autour des minimums locaux.

2.4.2 Le programme Rosetta fixbb

Rosetta est une suite de logiciels spécialisés dans la modélisation moléculaire, initialement développée au sein du laboratoire de David Baker. Cette suite est aujourd'hui maintenue par plus de 150 développeurs de 23 universités et laboratoires différents. Rosetta balaie actuellement un large champ d'applications comprenant la prédiction de structure, l'amarrage moléculaire, le dessin de protéines ou le raffinement des structures de membranes. Dans le domaine du dessin de protéines, Rosetta propose plusieurs outils prenant en compte ou non la flexibilité du squelette de la protéine ou encore spécialisés dans la modification des sites actifs d'enzymes. Par la suite nous utiliserons le programme fixbb qui utilise un squelette fixe et une bibliothèque de rotamères. Il est donc très proche du fonctionnement de Proteus. La bibliothèque de rotamères utilisée par fixbb, Dunbrack 2010, est une bibliothèque squelette-dépendante (Shapovalov & Dunbrack [2011]).

2.4.2.1 Fonction d'énergie

La fonction d'énergie utilisée par Rosetta est une fonction dite statistique (Park *et al.* [2016] ; Alford *et al.* [2017]). Elle repose sur dix-huit termes énergétiques pondérés par des constantes, certains correspondant à des forces physiques, et d'autres à des termes statistiques (Andrew Leaver-Fay *et al.* [2013]). Parmi les termes physiques, on retrouve un terme décrivant les interactions de van der Waals, modélisées par un potentiel de Lennard-Jones, un terme électrostatique modélisé par un terme de Coulomb dont la constante diélectrique est dépendante de la distance entre les atomes et un terme qui modélise les liaisons hydrogènes. L'énergie de

solvatation est modélisée par le modèle de Lazaridis-Karplus (Lazaridis & Karplus [1999]). À ces termes physiques s'ajoutent des termes statistiques issus de l'étude de structures cristallographiques. Ces termes correspondent par exemple à la probabilité d'observer un rotamère pour un squelette donné ou encore à une énergie interne propre à chaque rotamère.

2.4.2.2 Algorithme d'exploration

La méthode d'exploration utilisée par Rosetta correspond à une exploration par Monte Carlo en recuit simulé (Kirkpatrick *et al.* [1983]). Le principe consiste à effectuer une simulation à haute température puis à diminuer progressivement la température du système. À haute température, le système peut explorer des conformations de hautes énergies et passer des barrières énergétiques élevées. Lorsque la température diminue, les états de plus faibles énergies deviennent plus probables. Lorsque le système atteint zéro degré le système occupe théoriquement l'état de plus basse énergie. En pratique, ce type d'approche ne permet pas d'obtenir le minimum global en raison de la présence de nombreux minimums locaux et de la taille de l'espace à explorer.

Chapitre 3

Paramétrisation du modèle pour le dessin des domaines PDZ

Dans ce chapitre nous décrivons la paramétrisation du modèle déplié de la protéine à travers l'optimisation des énergies de référence. En effet, la description de l'état déplié est indispensable pour calculer la stabilité des séquences produites. Certains programmes comme Rosetta possèdent des énergies de référence suffisamment générales pour pouvoir être appliquées à toutes les protéines. Au contraire, Proteus nécessite une étape d'optimisation afin de reproduire au mieux la composition en acides aminés de la protéine étudiée. Nous allons tout d'abord décrire la procédure d'optimisation. Dans un second temps, le nouveau jeu de paramètres sera testé en redessinant entièrement les séquences de quatre domaines PDZ. Elles seront ensuite comparées aux séquences générées par Rosetta et aux séquences naturelles des domaines PDZ.

3.1 Optimisation des énergies de référence

3.1.1 Protocole d'optimisation

Nous allons optimiser les énergies de référence pour reproduire les fréquences en acides aminés observées dans les domaines PDZ. Cette étape est extrêmement importante puisqu'un mauvais calibrage des énergies de référence pourrait entrainer une sur abondance de certains types d'acides aminés. Comme nous l'avons vu dans le chapitre 2, un moyen de déterminer les énergies de référence consiste à calculer l'énergie de chaque acide aminé dans l'état déplié, par exemple avec un modèle MM/GBSA. Cette méthode, n'est cependant pas suffisante pour reproduire les fréquences expérimentales lors de l'exploration de l'espace des séquences. Une

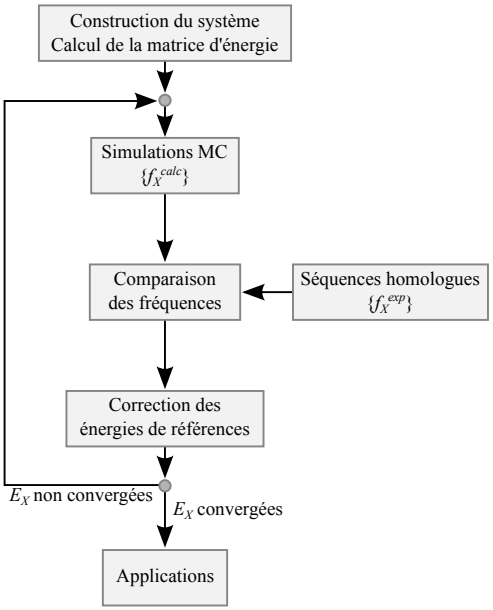


Figure 3.1 – Représentation schématique du protocole d'optimisation des énergies de référence.

phase d'optimisation empirique est donc nécessaire. Cette phase consiste à ajouter à chaque énergie de référence une correction empirique. Les fréquences expérimentales sont calculées à partir de séquences homologues à la protéine étudiée. Des simulations Monte Carlo sont effectuées de manière itérative en corrigeant à chaque étape les énergies de référence jusqu'à obtenir des fréquences proches des séquences expérimentales. Un schéma du déroulement de l'optimisation est présenté en figure 3.1. Nous allons maintenant détailler les différentes étapes.

3.1.2 Séquences expérimentales et modèles structuraux

3.1.2.1 Choix des modèles structuraux

Bien qu'homologues, les domaines PDZ n'ont pas exactement la même structure tridimensionnelle. L'utilisation d'un squelette fixe unique peut alors introduire un biais et un manque de transférabilité. Deux modèles structuraux sont donc pris en compte lors de l'optimisation. Ils correspondent aux domaines PDZ de Tiam1 et Cask dont les identifiants PDB sont respectivement 4GVD et 1KWA. Tous deux sont des domaines de classe II, c'est-à-dire qu'ils reconnaissent préférentiellement le motif Φ-X-Φ à l'extrémité C-terminale du ligand, Φ étant un acide aminé hydrophobe.

La structure 4GVD de Tiam1 correspond au domaine PDZ lié à son ligand naturel, le peptide Syndecan1 (Sdc1). La majorité des tests effectués par la suite étant faite sur la forme

apo du domaine, le peptide est retiré du modèle structural. La valeur du RMSD entre les structures cristallographiques de Tiam1 apo et holo (liée à un peptide) étant de seulement 0,5 Å, le retrait du peptide ne devrait pas nuire à la transférabilité du modèle entre ces deux formes. Bien que cristallisée sans ligand, la structure de Cask (1KWA) montre que l'extrémité C-terminale d'un autre domaine PDZ de la maille occupe le sillon de liaison du peptide. Comme pour Tiam1, les formes apo et holo de Cask sont probablement proches.

Afin de tester notre modèle par validation croisée, les seconds domaines PDZ de la Synténine (code PDB 1R6J) et de la protéine DLG2 (code PDB 2BYG) sont également sélectionnés. Ces domaines appartiennent à la classe I et reconnaissent le motif S|T-X-Φ à l'extrémité C-terminale du ligand. Comme pour Cask, ces deux structures n'ont pas été co-cristallisées avec un peptide mais le site de liaison est occupé par l'extrémité C-terminale d'un autre domaine PDZ de la maille.

Les quatres structures sont très proches, notamment au niveau de 14 positions du cœur hydrophobe (figure 3.2A). La plus grande dissimilarité se situe au niveau des boucles et des extrémités qui sont très flexibles. On peut également observer une légère rotation de l'hélice α_2 de Tiam1 qui n'est pas retrouvée dans les trois autres structures. Le RMSD entre les différentes protéines a des valeurs entre 1,0 et 2,1 Å. Le pourcentage d'identité entre les différentes séquences est compris entre 17 et 33%. Ainsi, les protéines Tiam1 et Cask possèdent une identité de séquences de 33% et un RMSD de 1,7 Å (calculé sur 42 C_α). La Synténine et DLG2 présentent le pourcentage d'identité le plus faible (17%) mais sont structurellement les plus proches avec un RMSD de 1,0 Å calculé sur 60 C_α (figure 3.2B).

Les séquences des quatre domaines sont également comparées à un sous-ensemble de l'alignement Pfam *seed* (figure 3.3). Les 14 positions de cœur sont bien conservées dans l'alignement. On observe néanmoins quelques arginines, lysines et glutamines à certaines de ces positions. Cela peut s'expliquer par la petite taille des domaines PDZ qui peut permettre aux longues chaines aliphatiques de l'arginine et de la lysine d'être enfouies tout en exposant leur tête polaire au solvant.

3.1.2.2 Recherche de séquences homologues proches

Les séquences de Tiam1 et Cask ne permettent pas de décrire la diversité des domaines PDZ. Il faut donc agrandir le jeu de séquences expérimentales tout en restant compatible avec les squelettes de Tiam1 et Cask. Le jeu de séquences est donc élargi en recherchant

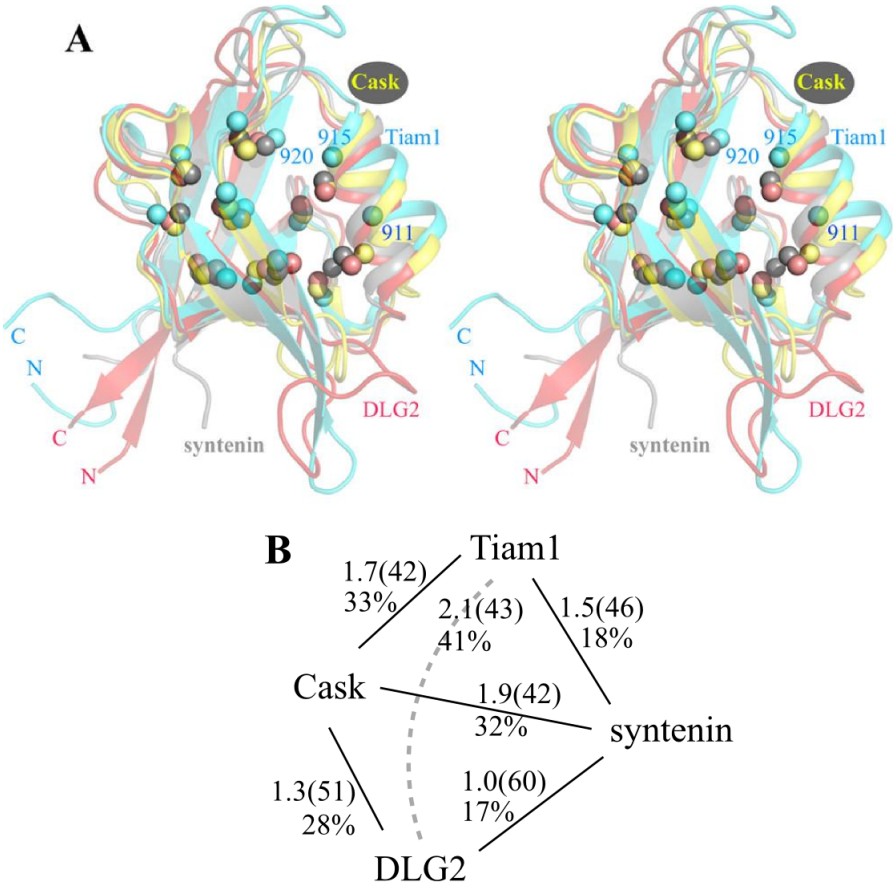


Figure 3.2 – Modèles structuraux utilisés pour l'optimisation des énergies de référence. A : Structure tridimensionnelle des 4 domaines PDZ. Les atomes C_β des 14 positions du cœur sont représentés par les sphères (les numéros de résidus correspondent à ceux de Tiam1). B : Proximité entre les domaines PDZ. Pour chaque lien, le pourcentage d'identité et le RMSD (Å) entre les atomes du squelette sont indiqués. Le nombre entre parenthèses correspond au nombre d'atomes C_α utilisés pour calculer le RMSD.

des séquences homologues de Tiam1 et Cask dans la base de données Uniprot. Pour cela le programme BLAST est utilisé (Altschul *et al.* [1990]) avec Tiam1 et Cask comme requêtes et la matrice de score Blosum62. Seules les séquences ayant un score d'identité entre 60% et 85% par rapport à la séquence requête sont conservées. Pour limiter la redondance, si deux séquences ont un pourcentage d'identité supérieur à 95%, seule l'une des deux est conservée. Nous obtenons ainsi 50 séquences homologues pour Tiam1 et 126 pour Cask. Ces ensembles seront notés $\mathcal{H}_T$ et $\mathcal{H}_C$ respectivement (Annexe A, figures A.1 et A.2). Les fréquences en acides aminés sont ensuite calculées pour chaque jeu de séquences $\mathcal{H}$ en moyennant sur toutes les séquences et toutes les positions.

La composition du cœur étant différente du reste de la protéine, avec une plus grande proportion d'acides aminés hydrophobes, les fréquences ont été calculées séparément pour les

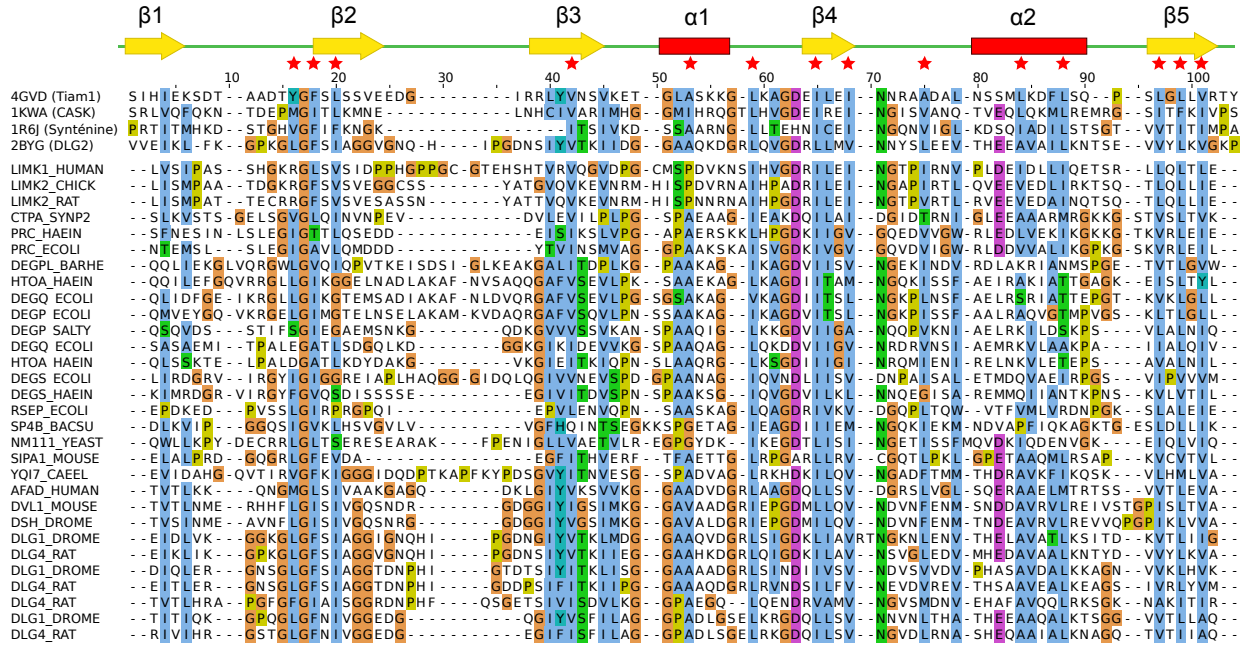


Figure 3.3 – Alignement des quatre séquences PDZ sélectionnées avec des séquences de l'alignement Pfam *seed*. Les quatre premières séquences sont les séquences testées. Les 30 autres sont issues de l'alignement Pfam *seed*. Les 14 positions du cœur hydrophobe sont indiquées par les étoiles rouges.

régions enfouies et exposées. Pour déterminer le caractère enfoui ou exposé d'un résidu, nous nous sommes basés sur leur surface relative exposée au solvant (RASA pour *Relative accessible surface area*). Les résidus ayant moins de 20% de leur surface exposée sont considérés comme enfouis. Ce seuil a été choisi de telle sorte qu'environ la moitié des positions des deux domaines PDZ sont considérées comme enfouies. Cette séparation des résidus en deux groupes, enfouis et exposés, prend en compte implicitement l'existence de structures résiduelles dans l'état déplié. Cela suppose que les résidus conservent une partie de leur propriété enfouie/exposée dans cet état. De plus, cette approche permet de rendre le modèle moins sensible aux variations de la longueur des boucles exposées et à la différence entre les proportions enfouie/exposée qui peuvent grandement varier entre homologues. Cela pourrait rendre le modèle plus facilement transférable à d'autres domaines PDZ.

Pour chaque protéine nous obtenons deux jeux de fréquences notées $\{f_t^b(\mathcal{H}), f_t^e(\mathcal{H})\}$, où t correspond au type d'acide aminé et les exposants b et e correspondent respectivement aux résidus enfouis (*buried*) et exposés (*exposed*). Une fois les fréquences calculées séparément pour les deux protéines, les valeurs obtenues sont moyennées comme suit, $f_t^b = (f_t^b(\mathcal{H}_T) + f_t^b(\mathcal{H}_C))/2$, pour chaque type et exposition.

La séparation des résidus en deux partitions double la taille du jeu d'énergies de référence à optimiser. Pour réduire le nombre de paramètres ajustables nous avons classé les acides aminés en groupes, basés sur leurs propriétés physico-chimiques. La composition initiale des groupes a été déterminée à partir de l'étude de Launay *et al.* [2007] par une approche de classification hiérarchique basée sur les scores de similarité Blosum50. Dans un second temps, certains groupes ont été scindés pour des raisons que nous préciserons plus loin.

3.1.3 Détermination des énergies de référence initiales

Pour déterminer les valeurs initiales des énergies de référence, deux méthodes différentes ont été utilisées. Elles mènent à des valeurs très proches.

3.1.3.1 Le peptide Sdc1 comme modèle déplié

Tout d'abord, nous avons calculé les énergies de référence en utilisant le ligand naturel de Tiam1, Sdc1, extrait du complexe Tiam1:Sdc1. Sdc1 se lie à Tiam1 en formant un feuillet β. Sa conformation modélise donc un état étendu dans lequel les chaines latérales interagissent peu entre elles. Des simulations de Monte Carlo où les cinq dernières positions du peptide pouvaient muter librement ont été effectuées. Pour chaque type échantillonné, la différence d'énergie par rapport à l'alanine (choisie arbitrairement comme référence) est calculée puis moyennée sur quatre positions, en excluant la dernière. Les valeurs obtenues constituent le jeu d'énergies de référence initiales.

3.1.3.2 Énergies diagonales

Une autre méthode, plus simple, consiste à déterminer les énergies de référence initiales à partir des énergies contenues dans la diagonale de la matrice d'énergie. En effet, la diagonale de la matrice décrit les interactions de chaque résidu avec lui-même et avec le squelette de la protéine. Les résidus exposés ayant peu d'interactions avec le squelette, leur état peut s'apparenter à un état déplié. Leur énergie diagonale correspond donc principalement à l'énergie interne au résidu et à l'énergie de solvatation. Ainsi, pour chaque position exposée, le rotamère de meilleure énergie est sélectionné pour chaque type d'acide aminé. Les énergies diagonales sont ensuite moyennées par type sur toutes les positions exposées. Les valeurs obtenues constituent le jeu d'énergies de référence initiales.

3.1.4 Optimisation des énergies de référence

La méthode d'optimisation consiste à modifier de manière itérative les valeurs des énergies de référence jusqu'à reproduire la composition en acides aminés expérimentale. Pour cela, à chaque itération, la fréquence de chaque type/groupe est comparée à la fréquence observée dans les séquences expérimentales. Une correction est ensuite appliquée de la manière suivante :

$$E_t^r(n+1) = E_t^r(n) - kT \ln \frac{\langle n(t) \rangle_n}{n_t^{exp}} \tag{3.1}$$

avec n le numéro de l'itération, $n_t^{exp} = N(t)/N$ la population moyenne de l'acide aminé t dans les séquences cibles, $\langle \rangle_n$ correspond à la moyenne d'un ensemble de séquences produites avec le jeu d'énergies de référence $E_t^r(n)$ et kT est une énergie thermique choisie de manière empirique égale à 0,5 kcal/mol. Si un acide aminé ou un groupe d'acides aminés est trop abondant dans les séquences produites, le terme correcteur va augmenter sa stabilité dans l'état déplié ce qui conduira à une réduction de sa présence dans les prochaines simulations Monte Carlo.

Cette méthode converge généralement en moins de 20 itérations. Il arrive cependant que les valeurs convergent lentement en fin d'optimisation et présentent un profil oscillatoire. Pour pallier ce problème nous modifions la règle d'optimisation dans les derniers pas en prenant en compte la correction appliquée au pas courant mais également celle appliquée au pas précédent, les deux valeurs étant respectivement pondérées par un poids de $\frac{2}{3}$ et $\frac{1}{3}$. Pour chaque itération, une simulation Monte Carlo avec échange de répliques de 500 millions de pas est effectuée avec huit répliques.

3.1.5 Optimisation dans un environnement natif

Lors de l'exploration Monte Carlo, la mutation d'une type vers un autre dépend de son environnement. Par exemple, il sera plus facile d'insérer un résidu hydrophobe dans un environnement hydrophobe. Pour faciliter la convergence des simulations nous imposons donc un environnement partiellement natif, proche de la composition naturelle des domaines PDZ. Pour ce faire, lors des simulations, une position sur deux seulement est autorisée à muter, l'autre pouvant uniquement explorer les rotamères associés à son type natif. Pour chaque système, deux ensembles d'acides aminés, ou «partitions», sont donc définis et deux simulations sont effectuées en parallèle. Seules les positions actives (qui peuvent muter) sont comptabili-

Tableau 3.1 – Pourcentages de résidus enfouis et exposés dans les deux partitions des protéines Tiam1 et Cask lors de l'optimisation.

Protéine	Part.	% exp.	% enf.
Tiam1	1	54	46
	2	33	67
Cask	1	62	38
	2	44	56

sées lors des calculs des fréquences théoriques. Les proportions des acides aminés enfouis et exposés dans les deux partitions sont indiquées dans le tableau 3.1.

3.2 Convergence des optimisations

3.2.1 Convergence des énergies de référence

20 itérations suffisent à obtenir la convergence des énergies de référence. Lors des dernières itérations les fluctuations des énergies de référence sont inférieures à 0,05 kcal/mol pour la plupart des résidus. Quelques résidus présentent néanmoins des fluctuations allant jusqu'à 0,1 kcal/mol, comme R, H et M enfouis ainsi que F et W exposés. Ces types sont les moins représentés et donc les plus sensibles aux fluctuations des énergies de référence. Les valeurs obtenues sont assez proches des valeurs obtenues pour le peptide étendu (tableau 3.2).

3.2.2 Convergence des fréquences

Le but de l'optimisation est de reproduire la composition des séquences expérimentales. Pour deux protocoles ($\epsilon_p = 8$ et $\epsilon_p = 4$), la composition théorique à la $20^{ème}$ itération au niveau des groupes est très proche de la composition expérimentale (tableau 3.3) avec un écart quadratique moyen de 1%, à la fois pour les résidus exposés et enfouis. L'analyses des compositions au sein des groupes montre une erreur légèrement plus importante avec un écart quadratique moyen pour le modèle $\epsilon_P = 8$ (respectivement $\epsilon_P = 4$) de 3,9% (2,1%) pour les résidus enfouis et 2,4% (2,4%) pour les résidus exposés. Dans quelques groupes, et notamment le groupe ACT enfouis, certains résidus sont surrepresentés (la cystéine dans ACT) au détriment des autres types. Les types au sein d'un groupe ne diffèrent que par les énergies physiques calculées sur le peptide étendu. Il semble donc que l'énergie physique présente un biais favorisant certains résidus. Pour limiter le problème, ces résidus ont été séparés du reste

Tableau 3.2 – Énergies de référence après optimisation. Les énergies ont été optimisées pour les constantes diélectriques $\epsilon_P = 4$ et 8. La colonne "Peptide" correspond aux valeurs initiales calculées à partir de Sdc1.

	$\epsilon_P = 8$			$\epsilon_P = 4$		
Residus	Peptide	Enfouis	Exposés	Peptide	Enfouis	Exposés
ALA	0,00	0,00	0,00	0,00	0,00	0,00
CYS	-0,85	-0,85	-0,85	-0,60	-0,60	-0,60
THR	-5,44	-5,44	-5,44	-8,22	-8,22	-8,22
SER	-6,43	-3,71	-4,74	-5,05	-5,68	-6,44
ASP	-17,28	-11,90	-15,88	-17,73	-20,31	-25,21
GLU	-17,35	-11,97	-15,95	-15,09	-17,67	-22,57
ASN	-12,25	-7,82	-10,22	-16,34	-17,70	-20,67
GLN	-11,50	-7,07	-9,47	-13,00	-14,35	-17,32
[a]HIS^+	9,02	12,53	9,73	15,04	13,52	9,80
[a]HIS_ϵ	6,98	10,49	7,69	11,42	9,90	6,18
[a]HIS_δ	7,35	10,86	8,06	12,14	10,62	6,89
ARG	-36,90	-32,00	-35,18	-48,42	-54,08	-57,90
LYS	-11,71	-6,76	-10,17	-7,08	-8,41	-12,35
ILE	4,22	4,63	3,63	1,67	5,30	4,00
VAL	-0,15	0,26	-0,74	-4,52	-0,89	-2,19
LEU	-0,53	-0,12	-1,12	-4,60	-0,97	-2,27
MET	-1,78	-2,05	-2,40	-2,55	-1,11	-2,17
PHE	-3,98	-0,23	-4,17	-1,21	0,66	-4,01
TRP	-5,96	-2,21	-6,15	-1,71	0,17	-4,50
TYR	-10,09	-5,80	-9,82	-7,37	-7,87	-12,52

[a]État de protonation de His.

du groupe. Les sept groupes initiaux ont ainsi été séparés en douze groupes. Cela n'a toutefois pas permis de résoudre entièrement le problème.

3.3 Validation des paramètres

Les énergies de référence optimisées permettent de retrouver des compositions proches de celles des séquences expérimentales. Ces résultats ont été obtenus à partir des simulations en environnement semi-natif (une partition active et une inactive). Pour tester le modèle produit, Tiam1 et Cask sont entièrement redessinées. Pour juger de la qualité des séquences par rapport à des séquences obtenues avec un autre programme de CPD, le programme fixbb de la suite Rosetta est également utilisé pour produire des séquences compatibles avec les squelettes de Tiam1 et Cask. Pour des raisons de clarté, nous ferons référence par la suite au programme fixbb par le terme Rosetta.

Tableau 3.3 – Composition en acides aminés (%) des séquences expérimentales et produites par Proteus.

type	Sequences Exp.				Proteus $\epsilon_P = 8$				Proteus $\epsilon_P = 4$			
	Enfoui		Exposé		Enfoui		Exposé		Enfoui		Exposé	
	type	groupe	type	groupe	type	groupe	type	groupe	type	groupe	type	groupe
A	5,9		4,6		4,1		7,2		8,1		5,6	
C	1,5	11,2	1,2	13,4	8,6	12,7 [1,5]	5,8	13,6 [0,2]	4,4	12,5 [1,3]	3,1	13,8 [0,4]
T	3,8		7,6		0,0		0,6		0,0		5,1	
S	4,7	4,7	10,2	10,2	4,9	4,9 [0,2]	10,7	10,7 [0,5]	4,4	4,4 [-0,3]	10,7	10,7 [0,5]
D	3,5	9,6	6,2	16,7	7,4	9,4	8,0	16,1	3,3	11,0	0,9	16,9
E	6,1		10,5		2,0	[-0,2]	8,1	[-0,6]	7,7	[1,4]	16,0	[0,2]
N	1,9	2,7	7,4	16,1	1,8	2,8	8,6	17,1	1,0	2,5	3,5	15,7
Q	0,8		8,7		1,0	[0,1]	8,5	[1,0]	1,5	[-0,2]	12,2	[-0,4]
H^+	0,7		4,7		0,1		1,8		0,5		3,9	
H_ϵ	0,0	0,7	0,0	4,7	0,6	0,9 [0,2]	2,2	4,5 [-0,2]	0,2	0,8 [0,1]	0,7	4,7 [0,0]
H_δ	0,0		0,0		0,2		0,5		0,1		0,1	
I	15,7		4,1		25,1		8,4		15,4		2,7	
V	13,5	49,6	5,5	14,4	12,8	46,7 [-2,9]	3,3	15,3 [0,9]	18,0	48,8 [-0,8]	2,1	14,4 [0,0]
L	20,4		4,8		8,8		3,6		15,4		9,6	
M	5,0	5,0	1,4	1,4	5,9	5,9 [0,9]	1,4	1,4 [0,0]	3,9	3,9 [-1,1]	1,4	1,4 [0,0]
K	6,5	6,5	10,1	10,1	5,5	5,5 [-1,0]	10,8	10,8 [0,7]	6,9	6,9 [0,4]	11,6	11,6 [1,5]
R	1,8	1,8	9,5	9,5	2,2	2,2 [0,4]	9,1	9,1 [-0,4]	1,5	1,5 [-0,3]	9,8	9,8 [0,3]
F	5,0	5,0	0,4	0,4	3,2	5,5	0,3	0,5	3,1	5,1	0,1	0,4
W	0,0		0,0		2,3	[0,5]	0,2	[0,1]	2,0	[0,1]	0,3	[0,0]
Y	2,9	2,9	0,9	0,9	3,4	3,4 [0,5]	0,9	0,9 [0,0]	2,6	2,6 [-0,3]	0,8	0,8 [-0,1]
G	0,0	0,3	1,7	2,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
P	0,3		0,4		0,0	[-0,3]	0,0	[-2,1]	0,0	[-0,3]	0,0	[-0,4]
	type	groupe	type	groupe	type	groupe	type	groupe	type	groupe	type	groupe

Les compositions sont données pour les positions enfouies et exposées soit par type soit par groupe. La valeur entre crochets correspond à l'écart par rapport aux séquences expérimentales

3.3.1 Génération des séquences Proteus

Des séquences ont été produites en utilisant les énergies de référence optimisées précédemment. Pour cela, des simulations de Monte Carlo de 750 millions de pas avec échanges de répliques sont effectuées. Les huit répliques ont des énergies thermiques kT comprises entre 0,125 et 3 kcal/mol. Toutes les positions sont autorisées à muter, exceptées les prolines et les glycines. Les simulations ont été effectuées avec une fonction d'énergie MMGBSA. À partir des trajectoires, les 10000 séquences de meilleures énergies, toutes répliques confondues, sont extraites et utilisées pour les analyses. Pour étudier l'impact de la température sur les

séquences explorées, les 10000 séquences de meilleures énergies des répliques comprises entre 0,263 et 0,888 (0,263, 0,395, 0,592 et 0.888) sont également extraites.

3.3.2 Génération des séquences Rosetta

Des séquences sont également produites à l'aide du programme Rosetta, largement utilisé pour le dessin de protéine. Les simulations ont été effectuées avec la version 2015.38.58158 en utilisant la ligne de commande :

```
fixbb -s Tiam1.pdb -resfile Tiam1.res -nstructu 10000 -ex1 -ex2 -linmem_ig 10
```

Les paramètres `ex1` et `ex2` permettent d'améliorer l'exploration des rotamères des résidus enfouis. L'option `linmem_ig` permet de calculer les interactions à la volée. Pour être directement comparables aux résultats de Proteus toutes les positions sont autorisées à muter exceptées les Pro et les Gly et 10000 séquences de basse énergie sont générées.

3.3.3 Caractérisation des séquences de Tiam1 et Cask

3.3.3.1 Compatibilité des séquences avec le repliement des domaines PDZ

La première étape de validation des séquences consiste à savoir si elles sont compatibles avec le pli des domaines PDZ. Pour cela, le programme de reconnaissance de plis *Superfamily* est utilisé (Gough *et al.* [2001] ; Wilson *et al.* [2007]). À l'aide de modèles de Markov cachés produits à partir de structures tridimensionnelles connues, *Superfamily* classe les séquences d'après la classification structurale des protéines SCOP (*Structurale Classification of Proteins*). Chaque séquence est ainsi attribuée à une superfamille et à une famille SCOP. Nous utilisons la version 1.75 de la base de données SCOP représentée par 15438 modèles et la version 3.5 du programme *Superfamily*. Pour chaque séquence, le programme renvoie les superfamilles et familles identifiées, chacune associée à une *E-value*, ainsi que la longueur de l'alignement. Les résultats obtenus sont présentés dans le tableau 3.4.

Les séquences produites par Proteus avec le modèle $\epsilon_P = 4$ sont toutes attribuées à la superfamille des *PDZ-like domains*. Pour Tiam1, 53% des séquences sont attribuées à la famille des domaines PDZ. Les 47% restants sont attribués aux familles des protéases à sérine *HtrA-like* (39%) et des protéases *tail specific* (8%). La superposition de la structure de Tiam1 avec deux représentants de ces familles (1Y8T:A pour les protéases à sérine *HtrA-like* et 1FC6:A

Tableau 3.4 – Reconnaissance de pli des séquences générées par Proteus et Rosetta. Pour chaque modèle, 10000 séquences ont été soumises à *Superfamily*.

Protéine	Modèle	[a]Match/long séquence	[b]E-value superfamille	[c] # succès superfamille	[b]E-value famille	[c] # succès famille
Tiam1	Proteus, ϵ_P=4	53/94	$1{,}0\times10^{-4}$	10000	$7{,}0\times10^{-2}$	5259
Cask	Proteus, ϵ_P=4	76/83	$5{,}1\times10^{-7}$	10000	$1{,}6\times10^{-2}$	10000
Tiam1	Proteus, ϵ_P=8	64/94	$1{,}2\times10^{-4}$	9920	$5{,}2\times10^{-2}$	9058
Cask	Proteus, ϵ_P=8	71/83	$3{,}2\times10^{-7}$	10000	$8{,}2\times10^{-3}$	10000
Tiam1	Rosetta	65/94	$4{,}4\times10^{-4}$	9035	$2{,}8\times10^{-2}$	9030
Cask	Rosetta	68/83	$2{,}8\times10^{-5}$	9832	$7{,}5\times10^{-3}$	9832

[a]Taille moyenne des séquences reconnues par Superfamily par rapport à la taille de la séquence. [b]Valeur moyenne de la *E-value* pour l'attribution à la bonne superfamille/famille SCOP. [c]Nombre de séquences (sur 10000) attribuées à la bonne superfamille/famille.

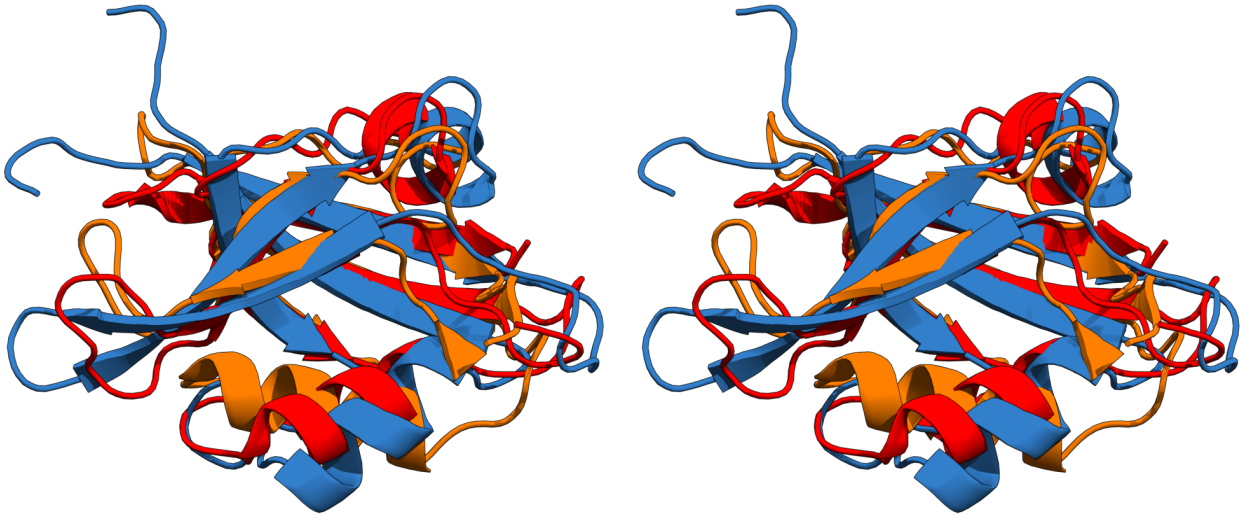


Figure 3.4 – Superposition de Tiam1 avec les représentants des familles *HtrA-like serin* et *Tail specific protease*. Tiam1 est en bleu. Les régions reconnues par *Superfamily* de la protéase à serine (1Y8T, chaine A) et de la protéase du photosystem II D1 (1FC6, chaine A) sont respectivement représentées en orange et en rouge.

pour les protéases *tail specific*) montre que le repliement des trois domaines est très proche, la principale différence se situant dans l'orientation de l'hélice α_2 (figure 3.4).

Le modèle utilisant une constante diélectrique $\epsilon_P = 8$ donne de meilleures performances pour le squelette de Tiam1 avec 99% des séquences attribuées à la superfamille des *PDZ-like domains* et 91% à la famille des domaines PDZ. Bien que plus de séquences soient reconnues avec ce modèle, les *E-values* obtenues pour Tiam1 restent plus élevées que celles de Cask.

Les séquences produites par Rosetta obtiennent des résultats similaires à Proteus avec 90% d'attribution à la bonne superfamille/famille pour Tiam1 et 98% pour Cask. Les *E-values*

associées aux séquences de Tiam1 sont similaires aux séquences Proteus tandis que celles pour Cask sont légèrement plus basses. Tout comme Proteus, les performances de Rosetta semblent légèrement moins bonnes pour le squelette de Tiam1.

3.3.3.2 Comparaison des séquences générées avec l'alignement Pfam

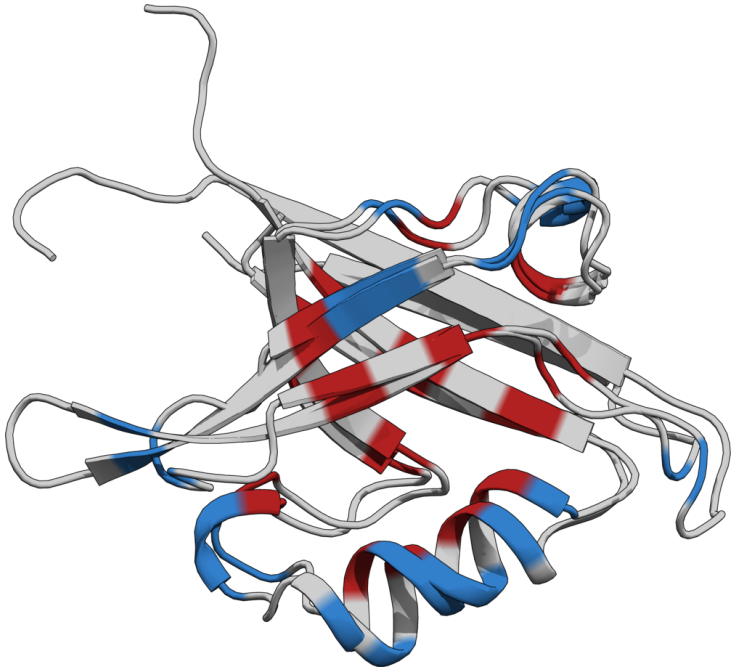
Les séquences produites ont également été comparées aux séquences natives et homologues, dont celles de la base de données Pfam (Sonnhammer *et al.* [1997]). Cette base de données rassemble des alignements de séquences des différentes familles de protéines. Pour ce faire, un petit jeu de séquences appelé *seed* (44 séquences dans le cas des domaines PDZ) est aligné manuellement. Puis toutes les séquences de la famille sont alignées automatiquement à l'aide de modèles de Markov cachés issus de l'alignement *seed*. L'alignement PDZ complet compte actuellement 38522 séquences.

Nous avons fait le choix de travailler avec le sous-ensemble RP55 contenant 12255 séquences pour réduire la taille du jeu de données. Le seuil de 55% permet de diminuer la redondance tout en conservant une bonne diversité (Chen *et al.* [2011]). Le score de similarité moyen de chaque séquence par rapport aux séquences Pfam est calculé en utilisant la matrice de score Blosum40 et une pénalité de *gap* de -6. Cette matrice permet de caractériser des homologues distants. Pour comparaison, le score de chaque séquence Pfam est également calculé par rapport à l'alignement Pfam. Pour ne pas biaiser les résultats, l'alignement d'une séquence Pfam contre elle-même n'est pas pris en compte.

Les scores de similarité sont calculés sur la protéine entière mais également sur 14 positions du cœur hydrophobe et 16 positions de surface déterminées par leur surface exposée au solvant. Les numéros et la position de ces résidus sont indiqués en figure 3.5. Ces deux ensembles sont également représentés sous forme de logos où chaque type est indiqué par son code à une lettre dont la taille est proportionnelle à sa fréquence (figures 3.6 et 3.7).

Les deux modèles Proteus $\epsilon_P = 4$ et $\epsilon_P = 8$ ont des performances comparables (figures 3.8 et 3.9). Les scores de similarité obtenus avec Proteus chevauchent la partie inférieure du pic des séquences Pfam. Les scores obtenus restent néanmoins inférieurs aux scores des séquences natives. Les résultats Rosetta sont très similaires à ceux de Proteus avec des performances légèrement meilleures dans le cas de Tiam1.

Les scores des positions de cœur sont meilleurs dans le cas de Proteus et se rapprochent du pic des séquences Pfam. Pour Cask, le score de similarité de certaines séquences égalise,



Pos. cœur		Pos. surface	
Tiam1	Cask	Tiam1	Cask
Y858	M501	A855	D498
F860	I503	E866	E509
L862	L505	R871	N511
V875	V515	N876	A516
A884	I524	S877	R517
L889	L530	K879	M519
I895	I536	E880	H520
I898	I539	K886	R526
A903	V544	K887	Q527
L911	L552	K890	H531
L915	L556	D904	A545
L920	I563	A905	N546
L922	F565	S909	E550
V924	I567	K912	Q553
		D913	K554
		S916	R557

A-Positions des résidus de cœur et de surface sur les structures de Tiam1 et Cask.

B-Numéros des résidus de cœur et de surface.

Figure 3.5 – Positions et numéros des résidus du cœur hydrophobe et de la surface des protéines Tiam1 et Cask. Les positions enfouies et exposées sont respectivement représentées en rouge et bleu.

voire dépasse, le score de la séquence native. Ces positions sont extrêmement contraintes géométriquement, ce qui limite les types possibles. Les logos (figure 3.6) montrent que les positions enfouies présentent la plupart du temps le type natif ou un type très présent dans l'alignement Pfam. Les deux principales différences se situent aux positions 898 et 903 de Tiam1. La présence d'une lysine dans le cœur à la position 898 peut paraître étonnante, mais en réalité la chaine aliphatique de la lysine traverse le cœur de la protéine pour exposer sa tête polaire au solvant entre le brin β_2 et l'hélice α_2. De même, la position 903 est proche de la surface de la protéine, ce qui permet à la sérine ou à l'asparagine d'exposer leur groupement polaire. Rosetta a, par ailleurs, tendance à introduire des résidus polaires aux positions 858, 862, 911 et 915 de Tiam1 ce qui pourrait expliquer ses moins bonnes performances pour les positions du cœur.

Les positions de surface présentent des scores négatifs à la fois pour les séquences Proteus et les séquences Rosetta mais restent proches des valeurs obtenues pour les séquences natives. La nature des résidus de surface est contrainte par les interactions de la protéine avec ses partenaires. Ces interactions n'étant pas prises en compte ici, les programmes de CPD ont tendance à y positionner des résidus polaires de manière aléatoire. Cela explique les faibles

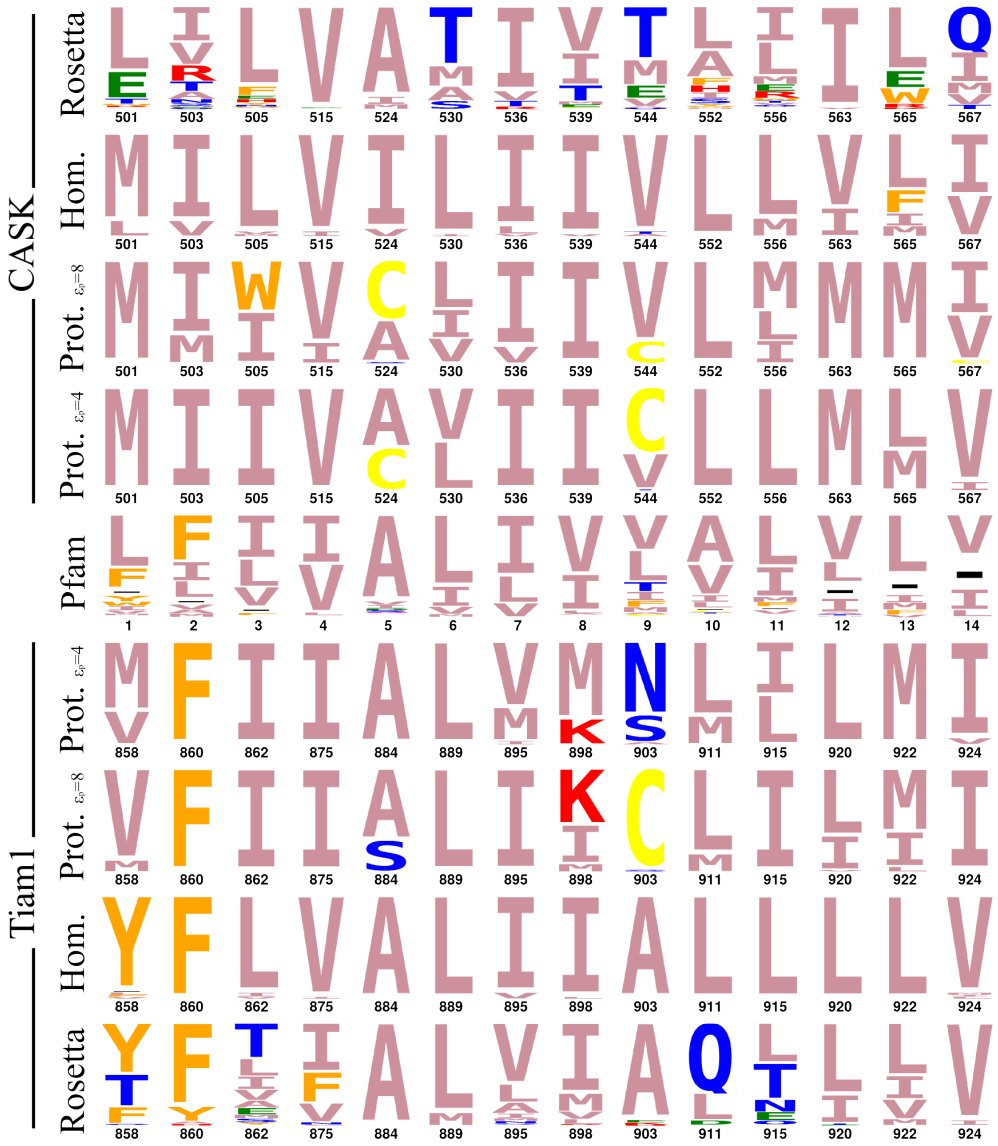


Figure 3.6 – Logos des positions de cœur des séquences de Tiam1 et Cask.

scores observés pour ces positions, qui sont par ailleurs très peu conservées dans l'alignement Pfam (figure 3.7).

Bien que les scores de similarité des séquences Proteus et Rosetta soient comparables, les scores d'identité par rapport à la séquence native sont plus élevés pour les séquences Rosetta (tableau 3.5). Ainsi, les séquences du modèle $\epsilon_P = 8$, en excluant (respectivement incluant) les Gly et Pro, sont de 20% (28%) pour Proteus contre 26% (34%) pour Rosetta, ce qui représente une différence d'environ 5 mutations pour Tiam1 et Cask.

Figure 3.7 – Logos des positions de surface des séquences de Tiam1 et Cask.

3.3.3.3 Entropie de séquence

Afin de comparer la diversité des séquences produites aux séquences naturelles, l'entropie des séquences par position est calculée de la manière suivante (Durbin *et al.* [2002]) :

$$S_i = -\sum_{j=1}^{6} f_j(i) \ln f_j(i) \tag{3.2}$$

où $f_j(i)$ est la fréquence du résidu j à la position i. Au lieu de calculer l'entropie pour les 20 types d'acides aminés, ces derniers sont rassemblés en six groupes : LIVMC, FYW, G, ASTP,

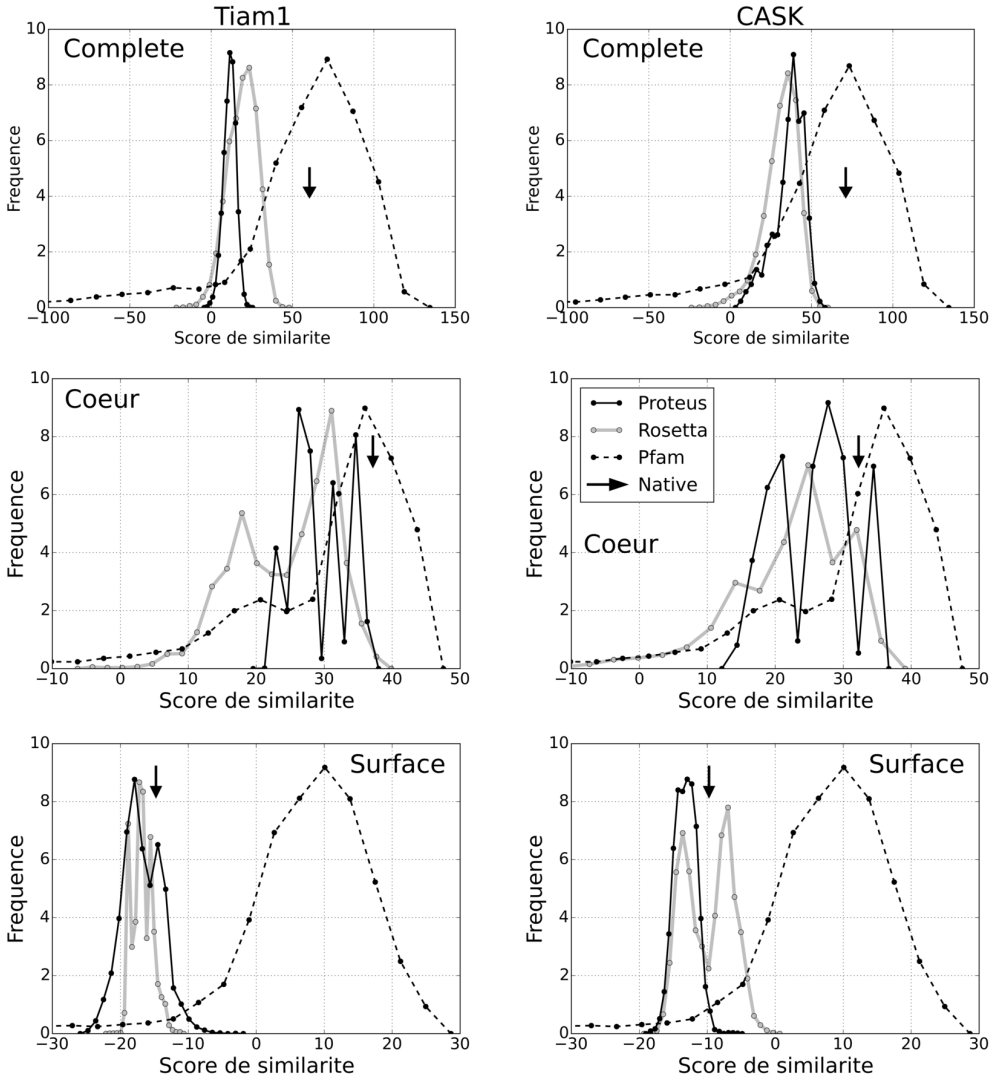


Figure 3.8 – Histogrammes des scores de similarités des séquences PDZ de Proteus (modèle $\epsilon_P = 8$) et Rosetta. Les scores ont été calculés par rapport à l'alignement Pfam RP55 en utilisant la matrice de score BLOSUM40. Les scores ont été calculés sur toute la protéine (haut), sur 14 résidus du cœur (milieu) et sur 16 positions de la surface (bas) pour les protéines Tiam1 (gauche) et Cask (droite). Le score de la séquence native est indiqué par une flèche verticale.

EDNQ et KRH. Ces groupes ont été définis à partir de la matrice Blosum50 et des énergies de contact entre résidus au sein des protéines (Murphy *et al.* [2000] ; Launay *et al.* [2007]). En prenant l'exponentielle de l'entropie, le nombre de type d'acides aminés qui apparaissent à chaque position peut être estimé. Ainsi, une valeur de deux pour une position donnée signifie que des acides aminés appartenant à deux des six groupes sont présents à cette position dans les séquences étudiées. Les valeurs obtenues sont ensuite moyennées sur les positions d'intérêt, c'est-à-dire la protéine complète, les positions de cœur ou de surface.

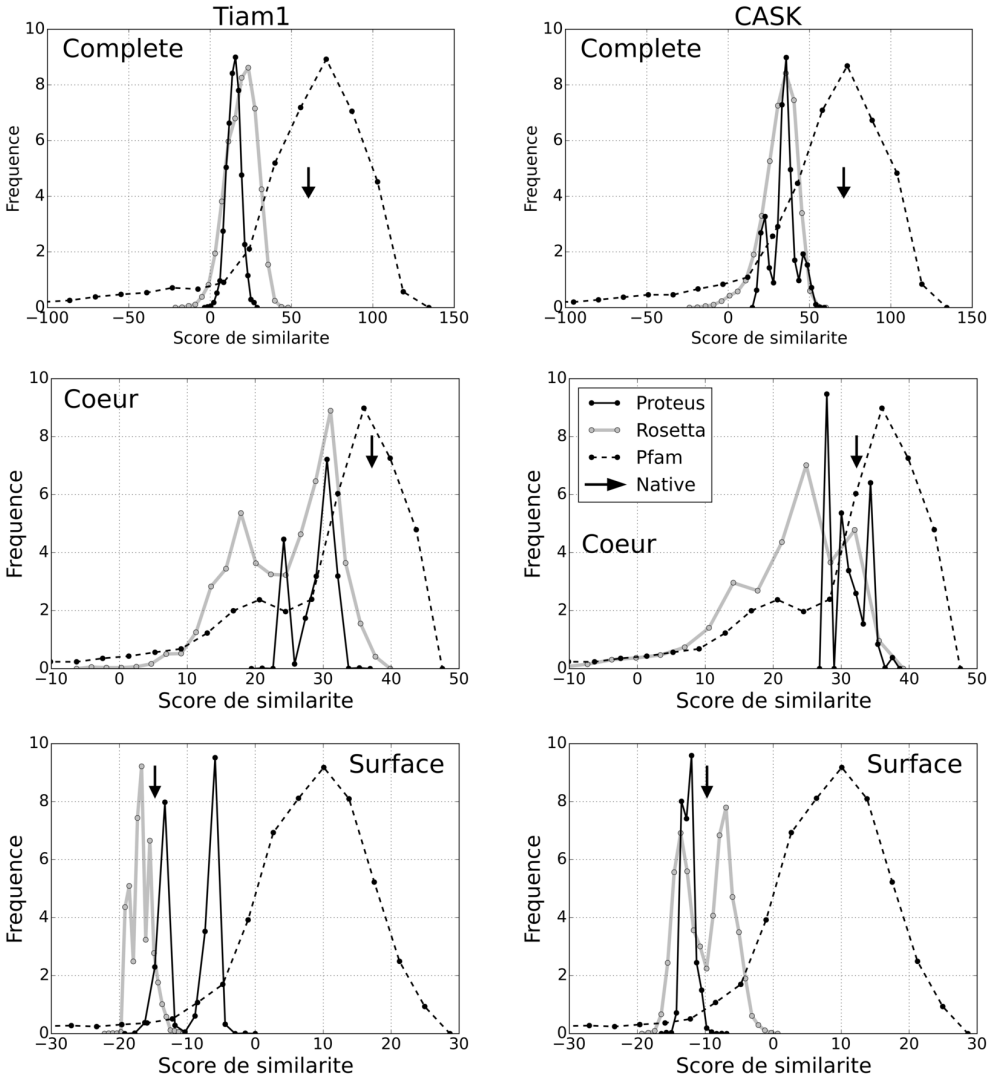


Figure 3.9 – Histogrammes des scores de similarités des séquences PDZ de Proteus (modèle $\epsilon_P = 4$) et Rosetta. Les scores ont été calculés par rapport à l'alignement Pfam RP55 en utilisant la matrice de score BLOSUM40. Les scores ont été calculés sur toute la protéine (haut), sur 14 résidus du cœur (milieu) et sur 16 positions de la surface (bas) pour les protéines Tiam1 (gauche) et Cask (droite). Le score de la séquence native est indiqué par une flèche verticale.

L'entropie moyenne des séquences Pfam est de 3,4 (tableau 3.6). Pris séparément, les squelettes de Tiam1 et Cask donnent des entropies plus basses, aussi bien avec Proteus qu'avec Rosetta. En regroupant les séquences de Tiam1 et Cask, l'entropie obtenue est de 2,2 pour Rosetta et comprise entre 1,8 et 1,9 pour Proteus. Cela montre que les squelettes de Tiam1 et Cask ne permettent pas d'explorer toute la diversité des séquences contenues dans l'alignement RP55. Les séquences issues de la réplique à température ambiante ($kT = 0{,}592$) possèdent une entropie plus élevée égale à 2,9.

Tableau 3.5 – Scores de similarité et pourcentages d'identité des séquences générées par Proteus et Rosetta. Les scores de similarité ont été calculés par rapport à un sous-ensemble de l'alignement Pfam dans domaines PDZ (S_{Pfam}), aux séquences des homologues proches ($S_{hom.}$) et de la séquence native (S_{native}) en utilisant une matrice de score BLOSUM40. Le score d'identité par rapport à la séquence native en prenant en compte les Gly et Pro a également été calculé (%ID$_{native}$).

Part.	Modèle	Prot.	S_{Pfam}	$S_{hom.}$	S_{native}	%ID$_{native}$
Prot, complète	Proteus $\epsilon_P = 4$	Tiam1	15,5 (4,2)	114,4 (9,6)	122,7 (110,0)	27,3 (2,3)
		Cask	35,2 (7,8)	138,3 (20,9)	143,5 (19,9)	29,5 (2,8)
	Proteus $\epsilon_P = 8$	Tiam1	12,3 (3,9)	101,5 (10,6)	111,7 (10,9)	25,4 (2,0)
		Cask	37,9 (9,4)	163,6 (19,3)	173,8 (18,8)	32,7 (2,7)
	Rosetta	Tiam1	21,2 (9,0)	146,6 (15,6)	173,1 (16,3)	35,6 (2,9)
		Cask	33,8 (10,5)	147,3 (17,3)	172,1 (19,2)	35,6 (3,1)
Cœur	Proteus $\epsilon_P = 4$	Tiam1	29,4 (2,9)	51,5 (6,7)	51,3 (6,8)	38,0 (6,7)
		Cask	31,3 (2,9)	53,3 (3,1)	52,5 (3,3)	56,4 (4,7)
	Proteus $\epsilon_P = 8$	Tiam1	30,1 (3,9)	47,8 (4,4)	47,6 (4,5)	40,0 (5,3)
		Cask	26,6 (5,4)	49,7 (5,5)	48,6 (6,0)	53,1 (8,5)
	Rosetta	Tiam1	25,4 (7,4)	56,3 (12,5)	56,8 (12,8)	59,7 (13,0)
		Cask	23,9 (8,7)	39,4 (12,9)	39,9 (12,9)	43,2 (12,0)
Surface	Proteus $\epsilon_P = 4$	Tiam1	-8,9 (3,7)	-18,0 (2,7)	11,9 (3,9)	18,4 (4,3)
		Cask	-12,1 (0,9)	11,8 (5,4)	13,1 (5,4)	9,7 (4,1)
	Proteus $\epsilon_P = 8$	Tiam1	-16,2 (2,9)	-19,3 (4,9)	16,6 (7,4)	19,8 (6,6)
		Cask	-12,9 (1,6)	16,3 (6,2)	13,9 (6,6)	8,3 (5,6)
	Rosetta	Tiam1	-16,5 (1,5)	-16,5 (5,9)	11,2 (5,2)	14,6 (4,2)
		Cask	-9,4 (3,5)	23,8 (7,2)	36,0 (8,8)	32,7 (7,4)

Tableau 3.6 – Entropie des séquences expérimentales et produites par Proteus et Rosetta.

ϵ_P	Part.		Cask		Tiam1		Cask+Tiam1	
		Pfam	Rosetta	Proteus	Rosetta	Proteus	Rosetta	Proteus
4	Complète	3.40	1.75	1.23	1.55	1.22	2.24	1.82
	Cœur	1.79	1.86	1.07	1.55	1.11	2.11	1.27
	Surface	4.33	1.73	1.34	1.64	1.25	2.49	1.95
8	Complète	3.40	1.75	1.19	1.55	1.52	2.24	1.93
	Cœur	1.79	1.86	1.20	1.55	1.13	2.11	1.42
	Surface	4.33	1.73	1.23	1.64	1.99	2.49	2.23

Les positions de cœur des séquences Rosetta ont une entropie moyenne similaire au reste de la protéine et proche de Pfam. Ce n'est pas le cas des séquences Proteus qui présente des valeur d'entropie plus faible pour ces positions.

3.3.4 Application du modèle à deux autres domaines PDZ

Les énergies de référence ont été optimisées sur les modèles structuraux de Tiam1 et Cask en utilisant des séquences d'homologues proches. Il reste à savoir si le modèle est transférable à d'autres domaines PDZ. Pour cela nous redessinons les domaines PDZ des protéines DLG2

Tableau 3.7 – Reconnaissance de pli des séquences produites en validation croisée. Pour chaque modèle, 10000 séquences ont été sélectionnées puis soumises au programme *Superfamily*.

Protéine	Modèle	[a]Match/long séquence	[b]E-value superfamille	[c] # succès superfamille	[b]E-value famille	[c] # succès famille
syntenin	Proteus, ϵ_P=8	69/91	1.3×10^{-2}	9999	4.0×10^{-3}	9999
DLG2	Proteus, ϵ_P=8	85/97	8.0×10^{-9}	10000	5.0×10^{-3}	10000
syntenin	Rosetta	76/82	7.3×10^{-13}	10000	1.8×10^{-3}	10000
DLG2	Rosetta	86/97	1.3×10^{-9}	10000	9.6×10^{-4}	10000

[a]Taille moyenne des séquences reconnues par Superfamily par rapport à la taille de la séquence. [b]Valeur moyenne de la *E-value* pour l'attribution à la bonne superfamille/famille SCOP. [c]Nombre de séquences (sur 10000) attribuées à la bonne superfamille/famille.

et de la Synténine. Cette étape a été effectuée avec le modèle optimisé pour une constante diélectrique $\epsilon_P = 8$. Les résultats *Superfamily* obtenus sont comparables à ceux obtenus sur Tiam1 et Cask avec 100% de bonne attribution pour la protéine DLG2 et 99.99% pour la Synténine (tableau 3.7). Ces performances sont similaires à celles de Rosetta avec toutefois une *E-value* associée à la superfamille beaucoup plus faible pour les séquences de la Synténine produites par Rosetta (mais des valeurs proches pour les *E-values* associées à la famille). Malgré les bons résultats *Superfamily*, les scores de similarités par rapport à Pfam obtenus avec Proteus sont plus faibles que ceux de Rosetta (figure 3.10). Les cœurs des deux protéines présentent des scores de similarité proches du pic de l'alignement Pfam bien que Rosetta donne de meilleurs scores.

Les performances obtenues lors de la validation croisée sont donc légèrement dégradées par rapport aux performances obtenues sur Tiam1 et Cask. Le jeu d'énergies de référence n'est donc pas parfaitement transférable. Il est de ce fait préférable d'optimiser spécifiquement les énergies de référence dans le cas de l'étude d'un autre domaine PDZ.

3.4 Introduction d'un potentiel de biais

Jusqu'à présent, les énergies de référence ont été optimisées de manière à reproduire les compositions observées dans les séquences homologues proches et ce indépendamment de la position. Il est également possible de contraindre chaque position à explorer des types proches des types observés dans un jeu de séquences expérimentales en pénalisant les séquences ayant peu de similarité avec les séquences de référence. Pour cela un biais énergétique est introduit

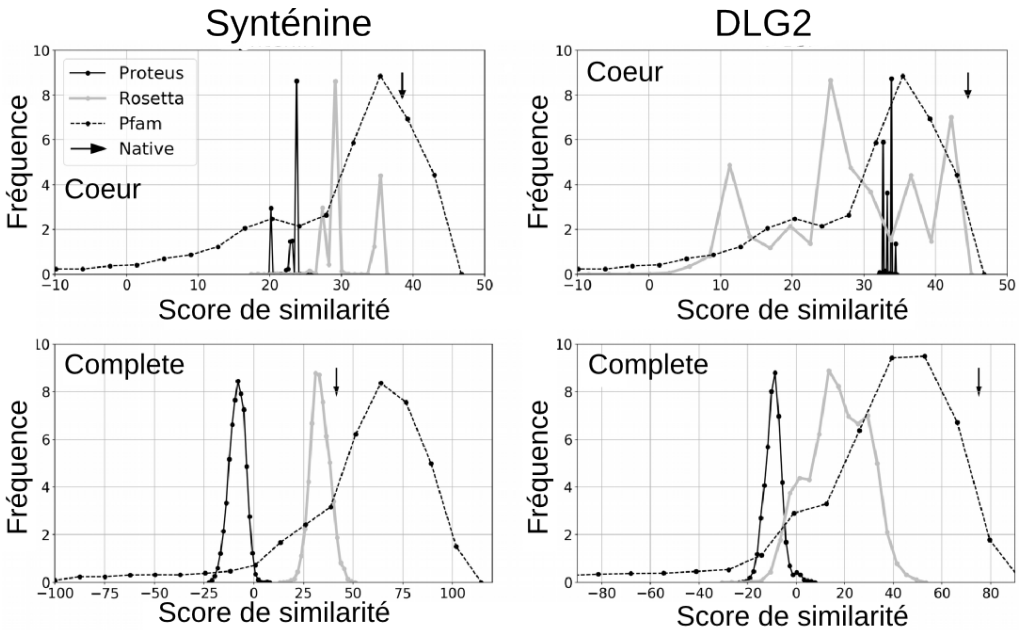


Figure 3.10 – Histogrammes des scores de similarités des séquences PDZ de Proteus (modèle $\epsilon_P = 8$) et Rosetta en validation croisée. Les scores de similarité ont été calculés par rapport à l'alignement Pfam RP55 en utilisant la matrice de score BLOSUM40. Les scores ont été calculés sur toute la protéine (bas) et sur 14 résidus du cœur (milieu) pour les protéines Synténine (gauche) et DLG2 (droite). Le score de la séquence native est indiqué par une flèche verticale.

pour chaque type et chaque position de la manière suivante :

$$\delta E_{biais} = c \sum_i (S_i^{rand} - S(t_i)) \tag{3.3}$$

où t_i correspond au type d'acide aminé à la position i, $S(t_i)$ représente le score de similarité Blosum40 par rapport à la position correspondante dans l'alignement Pfam. S_i^{rand} est le score moyen par rapport à la même colonne d'un type choisi aléatoirement de manière équiprobable. La valeur de c pondère le biais. Quatre valeurs de c ont été testées (0,1 ; 0,5 ; 1.0 ; et 2.0). Cette approche a été testée sur les protéines Tiam1 et Cask en utilisant une constante diélectrique $\epsilon_p = 8$. Les 10000 meilleures séquences dessinées sont testées à l'aide du programme *Superfamily* et comparées aux séquences de l'alignement Pfam.

L'introduction du biais Pfam améliore les résultats $Superfamily$ et ce dès une valeur de $c = 0{,}5$ puisque la quasi-totalité des séquences de Tiam1 et Cask sont reconnues comme appartenant à la famille des domaines PDZ (99,98% et 100% respectivement). De plus, la longueur moyenne des séquences reconnues est plus importante et les *E-values* associées à la superfamille sont 10^{-10} fois plus faibles (tableau 3.8). Les scores de similarité par rapport à

Tableau 3.8 – Reconnaissance de pli des séquences générées par Proteus en appliquant un potentiel de biais vers les séquences Pfam Pour chaque modèle, 10000 séquences ont été sélectionnées puis soumises au programme *Superfamily*.

Protein	c	[a]Match/seq length	[b]Superfamily E-value	[c]Superfamily success #	[b]Family E-value	[c]Family success #
Tiam1	0.1	61/94	1.57×10^{-3}	9980	4.45×10^{-2}	9174
Cask		77/83	8.86×10^{-12}	10000	9.76×10^{-3}	10000
Tiam1	0.5	77/94	5.60×10^{-16}	10000	2.76×10^{-2}	9998
Cask		78/83	8.69×10^{-19}	10000	7.37×10^{-3}	10000
Tiam1	1.0	78/94	8.48×10^{-19}	10000	2.75×10^{-2}	10000
Cask		78/83	9.11×10^{-23}	10000	5.23×10^{-3}	10000
Tiam1	2.0	76/94	1.99×10^{-22}	10000	8.36×10^{-2}	9988
Cask		77/83	4.32×10^{-26}	10000	4.53×10^{-3}	10000

[a]Taille moyenne des séquences reconnues par Superfamily par rapport à la taille de la séquence. length. [b]Valeur moyenne de la *E-value* obtnue pour l'attribution à la bonne superfamille/famille SCOP. [c]Nombre de séquences testées (10000 au total) attribuées à la bonne superfamille/famille.

Pfam indiquent également une très nette amélioration, puisqu'à partir de $c = 0{,}5$ les scores obtenus sont supérieurs au pic des scores des séquences Pfam aussi bien sur la protéine complète que sur les positions de cœur. Les positions de surfaces gagnent quelques points mais restent environ 20 points sous la valeur du pic des séquences Pfam (figure 3.11). Une valeur c plus importante pourrait leur être appliquée.

Les entropies calculées pour les différentes valeurs de c sont similaires à celles obtenues sans contrainte. Ainsi, malgré le biais appliqué, les séquences conservent leur diversité. Cette diversité reste toutefois plus faible que celle des séquences expérimentales.

Grâce à un biais dans les énergies de référence, il est donc possible de se rapprocher fortement des séquences Pfam. Dans le cas présent, l'ensemble des positions ont été contraintes mais il est possible de ne contraindre qu'une partie de la protéine ce qui pourrait s'avérer utile dans le dessin de sites actifs par exemple.

3.5 Dessin de positions impliquées dans la spécificité

Quatre positions de l'interface protéine:peptide responsables de la spécificité des domaines PDZ de Tiam1 et Tiam2 ont été explorées. Les protéines Tiam1 et Tiam2 sont deux protéines homologues dont les domaines PDZ présentent 28% d'identité de séquence (Shepherd *et al.* [2011]). Certains peptides sont reconnus par les deux protéines comme Caspr4 avec des valeurs

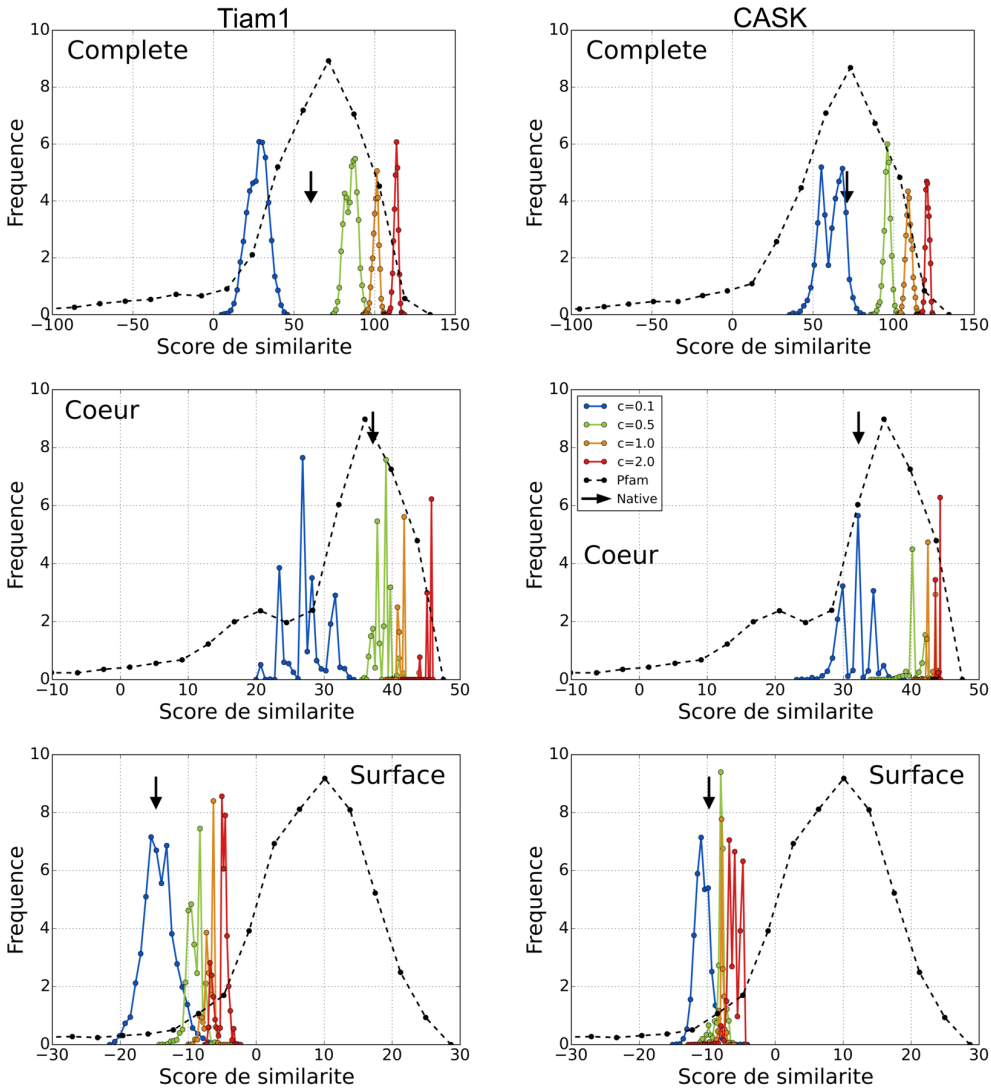


Figure 3.11 – Histogrammes des scores de similarités des séquences PDZ de Proteus en introduisant un biais vers les séquences Pfam Les scores ont été calculés par rapport à l'alignement Pfam RP55 en utilisant la matrice de score BLOSUM40 pour différents coefficients c. Les scores ont été calculés sur toute la protéine (haut), sur 14 résidus du cœur (milieu) et sur 16 positions de la surface (bas) pour les protéines Tiam1 (gauche) et Cask (droite). Le score de la séquence native est indiqué par une flèche verticale.

de K_d de 19,0 μM pour Tiam1 et 3,4 μM pour Tiam2. Malgré ce chevauchement, la reconnaissance d'autres partenaires reste très spécifique puisque Tiam1 est le seul capable de lier le peptide Sdc1 tandis que la Neurexine (Neu) n'est reconnue que par Tiam2. La comparaison des séquences des deux homologues a mis en évidence le rôle de quatre positions (911, 912, 915 et 920) dans la spécificité des deux domaines PDZ. La mutation de ces quatre positions vers les types présents dans la séquence de Tiam2 (LKLL → MEFV) réduit d'un facteur cinq l'affinité de Tiam1 pour le peptide Sdc1 et augmente d'un facteur 50 l'affinité pour le peptide

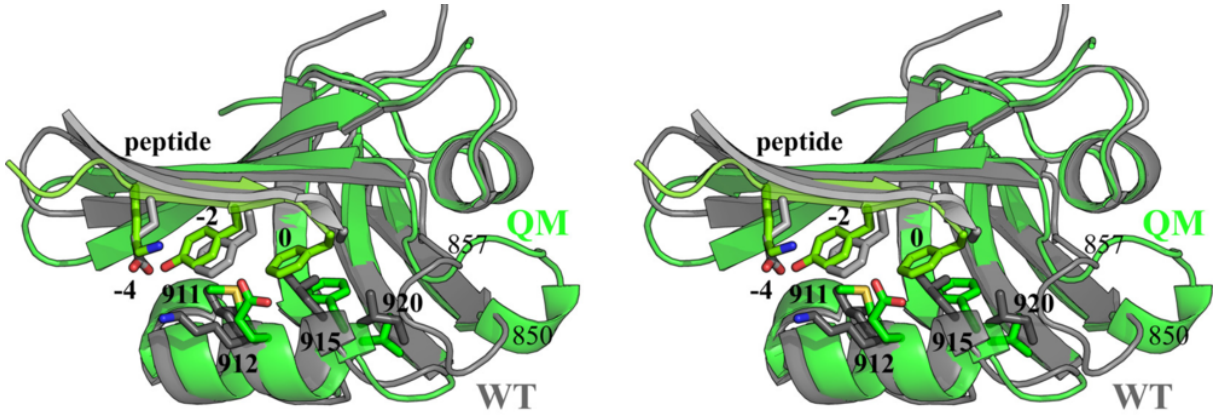


Figure 3.12 – Superposition des structures cristallographiques des complexes Tiam1:Sdc1 et QM:Caspr4. La structure de la séquence sauvage (LKLL, code PDB : 4GVD) est représentée en gris, tandis que la séquence du quadruple mutant (MEFV, code PDB : 4NXQ) est en vert.

Neu sans modifier l'affinité pour Caspr4. Le changement de spécificité observé pour le quadruple mutant (QM) s'explique principalement par une augmentation du volume de la poche S_0 et par l'existence d'interactions de type $\pi - \pi$, anion-π et S-π entre le peptide et les résidus F915, E912 et M911 (Liu *et al.* [2016]). Le dessin de ces positions permet de déterminer les performances du modèle dans le dessin d'interface protéine:peptide.

3.5.1 Modèles structuraux

Les squelettes des complexes Tiam1:Sdc1 (code PDB 4GVD) et QM:Caspr4 (code PDB 4NXQ) sont utilisés comme modèle pour le dessin des quatre positions de spécificité pour les formes apo et holo. Les deux structures sont très proches avec une valeur de RMSD de 0,9 Å. Les principales différences se situent au niveau de la boucle β_1-β_2 et de l'hélice α_2. Cette hélice est légèrement décalée dans la structure du quadruple mutant pour accommoder les chaines latérales des phénylalanines aux positions 915 et P_0 (figure 3.12). Cette différence laisse supposer que le squelette de la séquence sauvage décrira mieux les variants possédant une petite chaine latérale tandis que le squelette du mutant décrira mieux les grandes.

3.5.2 Génération des séquences

Au cours des simulations, les positions 911, 912, 915 et 920 peuvent muter vers tous les types exceptés Gly et Pro. Les autres positions sont autorisées à explorer les rotamères de leur type natif. Les modèles $\epsilon_P = 4$ et $\epsilon_P = 8$ ont donné des résultats similaires. Nous utiliserons

ici le modèle $\epsilon_P = 4$, qui donne de meilleures performances pour les positions enfouies du fait de sa constante diélectrique plus faible. Six systèmes sont simulés : WT et QM apo et liés à Sdc1 et Caspr4. Des simulations Monte Carlo de 2×10^8 pas avec 8 marcheurs sont effectuées. Les séquences issues du marcheur à température ambiante sont extraites et analysées par la suite.

3.5.3 Impact du squelette et du peptide sur les séquences générées

Les séquences obtenues sont comparées pour étudier l'effet du squelette et du peptide sur les séquences explorées (figure 3.13). Les six simulations permettent d'échantillonner la séquence native (LKLL) ou une séquence proche. Dans le cas du squelette de la séquence native (4GVD), L911 est majoritairement retrouvée lors de la simulation de la forme apo. Cette leucine est remplacée par les types homologues I et V lorsqu'un peptide est présent. Dans le cas du squelette du quadruple mutant, la position 911 échantillonne une gamme beaucoup plus large d'acides aminés, retrouvés pour la plupart dans l'alignement Pfam. L'ajout d'un peptide restreint l'espace exploré aux types I, L et M.

La plupart des simulations échantillonnent les types R, K et Q à la position 912 qui sont proches ou correspondent au type natif (K). La présence du peptide Caspr4 favorise l'échantillonnage du glutamate à la position 912, type présent chez le quadruple mutant. Le squelette du quadruple mutant semble favoriser la présence du glutamate y compris en l'absence de peptide.

Le type natif (L) est majoritaire à la position 915 avec le squelette WT dans les formes apo et liée au peptide Sdc1. Au contraire, le peptide Caspr4 semble imposer un résidu plus petit à cette position probablement à cause de la Phe à la position P_0. Avec le squelette du QM, le déplacement de l'hélice α_2 autorise l'insertion de résidus de plus grande taille. La présence du peptide Caspr4 ne permet cependant pas de retrouver de phénylalanine à la position 915 alors qu'elle est présente dans les deux autres simulations effectuées sur ce squelette. Enfin, pour la position 920, le squelette WT permet d'échantillonner le type natif (L) dans les trois simulations. La structure du quadruple mutant favorise au contraire le type retrouvé dans la séquence mutée (V). Les simulations sur les deux squelettes échantillonnent donc des séquences proches mais on constate tout de même un biais vers la séquence de chaque structure, probablement à cause du squelette fixe.

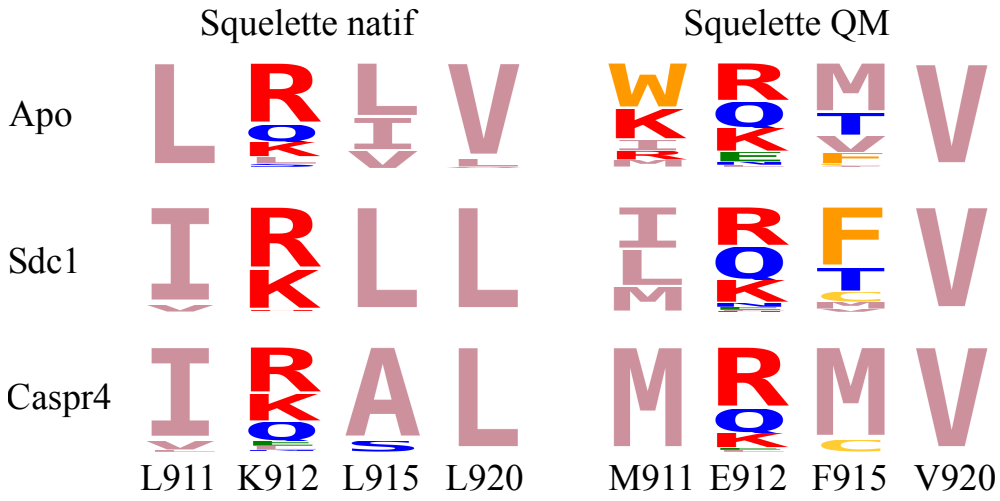


Figure 3.13 – Logos des séquencse obtenues lors du dessin des 4 positions de l'interface. L'exploration des séquences a été effectuée en présence ou en absence des peptides Sdc1 et Caspr4. Les modèles de squelettes sont issus de la séquence native (4GVD) ou du quadruple mutant (4NXR).

3.5.4 Stabilité des séquence générées

Les énergies Proteus donnent des informations sur la stabilité relative des séquences générées. Ainsi, en utilisant le squelette de la séquence sauvage dans sa forme apo, la simulation Monte Carlo retrouve la séquence native à seulement 2 kcal/mol de la séquence de meilleure énergie (KKLV). La séquence homologue LKML est la deuxième meilleure séquence et les homologues IKLL et LKLV ne sont qu'à 1-2 kcal/mol de la meilleure énergie. Le mutant MEFV n'est pas retrouvé, ni ses homologues proches. Cela est probablement dû à l'incapacité du squelette à accueillir une phénylalanine à la position 915. Lorsque le ligand Sdc1 est présent les résultats sont similaires. Les séquences LKLL, IKLL, VKLL et MKLL ont des énergies 1 à 2 kcal/mol au-dessus de la meilleure énergie. La séquence MKLL est intéressante car son affinité pour le peptide Sdc1 a été mesurée expérimentalement et est proche de la séquence sauvage.

Expérimentalement, les séquences native et mutante lient le peptide Caspr4 avec la même affinité et le quadruple mutant est légèrement moins stable (2 kcal/mol). Ces deux séquences devraient donc être échantillonnées aux cours des simulations du squelette natif en présence de Caspr4, ce qui n'est pas le cas. La séquence la plus proche de MEFV, IEAV, se situe à 2 kcal/mol de la meilleure séquence. Ces lacunes sont probablement dues encore une fois à

un conflit stérique entre la chaine P_0 de Caspr4 et la position 915 (L et F) pour le squelette sauvage.

En utilisant le squelette du quadruple mutant en l'absence de peptide, la séquence mutante (MEFV) est obtenue avec une énergie 5 kcal/mol au-dessus de la meilleure séquence (WYAM) et la séquence native LKLL est 2 kcal/mol plus haut. Les deux variants étant thermodynamiquement stables, la différence de 2 kcal/mol en faveur de la séquence mutante paraît raisonnable puisque l'exploration a été effectuée sur le squelette du quadruple mutant ce qui devrait favoriser sa séquence. En présence de Sdc1, la séquence MEFV est 6 kcal/mol au-dessus de la meilleure séquence IKLV, qui est l'homologue le plus proche de la séquence native. En présence du peptide Caspr4, la séquence mutante apparait à 7 kcal/mol par rapport à la meilleure séquence TKMV. Ses homologues MQMV et MEMV apparaissent à 5 kcal/mol. La séquences LKLL et ses homologues proches ne sont pas retrouvés, ce qui indique qu'elles présentent des énergies élevées.

3.5.5 Estimation de l'énergie libre de liaison

Les résultats précédents donnent des informations qualitatives sur l'effet des quatre positions étudiées. Il est également possible d'obtenir des informations plus détaillées en estimant l'affinité des complexes. Malgré la simplicité du modèle (squelette fixe et solvant implicite), cette approche peut constituer une première étape vers l'identification de mutations intéressantes.

Les analyses expérimentales ont montré que (1) les changements d'affinité liés aux mutations étaient indépendants puisque les énergies libres de couplage mesurées ne dépassent pas 0,4 kcal/mol ; (2) les mutations des positions 911, 915 et 920 vers des types homologues ont peu d'effets sur l'affinité ; (3) la mutation L915E réduit l'affinité de 0,5 à 1 kcal/mol pour les deux peptides ; (4) la mutation L915F affecte la liaison de manière différente selon le résidu présent à la position 912 et le peptide (Shepherd *et al.* [2011]).

Nous allons tenter de retrouver ces observations à partir des résultats de Proteus en calculant l'énergie libre de liaison relative des différents variants. Cela n'est cependant possible que si les deux variants ont été échantillonnés dans les simulations des formes apo et holo. En effet, si une séquence S est échantillonnée dans les formes apo et holo, on peut calculer une énergie moyenne $\langle E_{holo}(S)\rangle$ et $\langle E_{apo}(S)\rangle$ dans chaque état en moyennant les énergies Proteus

sur l'ensemble des conformations. La différence d'énergie libre entre deux séquences S et S' est alors estimée de la manière suivante :

$$\Delta\Delta E(S,S') = (\langle E_{holo}(S')\rangle - \langle E_{apo}(S')\rangle) - (\langle E_{holo}(S)\rangle - \langle E_{apo}(S)\rangle) \tag{3.4}$$

Certaines séquences n'étant pas présentes à la fois dans les formes apo et holo, les calculs ont également été effectués en prenant en compte non plus une séquence mais un groupe de séquences d'homologues. Comme précédemment, les énergies des différents conformations sont moyennées afin de pouvoir calculer la valeur de $\Delta\Delta E(S,S')$.

À partir du squelette WT, en prenant en compte les séquences homologues NKNN (où N $\in$ I,L,V,M), les énergies moyennes des formes apo et liée à Sdc1 sont respectivement de 0,9 $\pm$ 0,6 kcal/mol et 1,1 $\pm$ 0,5 kcal/mol ce qui représente une énergie libre de liaison de 0,2 $\pm$ 0,8 kcal/mol. La mutation entre les acides aminés de type I, L,V et M semble peu changer l'affinité pour Sdc1 ce qui est cohérent avec les données expérimentales. Pour les mutations NKNN $\rightarrow$ NENN, le changement d'énergie libre obtenu est de 0,75 kcal/mol pour les deux peptides. Cela est très proche des valeurs expérimentales qui sont respectivement de 0,94 kcal/mol et 0,55 kcal/mol pour Sdc1 et Caspr4. De même, la mutation NKNN $\rightarrow$ NKFN réduit l'affinité de 0,5 kcal/mol pour les deux peptides contre 1,2 et 0,8 kcal/mol (Sdc1 et Caspr4 respectivement) expérimentalement. Seul l'effet de la mutation NENN $\rightarrow$ NEFN n'est pas prédit correctement par notre modèle. Il prédit une perte de 0,9 kcal/mol pour Caspr4, contre un gain de 0,5 kcal/mol expérimentalement. Enfin, les changements de spécificité entre les peptides Sdc1 et Caspr4 prédits par notre modèle sont en accord avec les valeurs expérimentales. Par exemple, la mutation MKFV $\rightarrow$ MEFV favorise la liaison de Caspr4 par rapport à Sdc1 de 0,2 kcal/mol contre 0,5 kcal/mol expérimentalement.

Les simulations permettent également d'étudier la corrélation entre les résidus aux différentes positions. Les figures 3.14 et 3.15 présentent les covariances des résidus adjacents au cours des simulations de Monte Carlo (911-912, 912-915 et 915-920). La position 911 explore plus de types avec le squelette du quadruple mutant en l'absence de peptide. Cela avait déjà été remarqué lors de l'analyse des logos. La position 912 est peu sensible à la présence et au type du peptide. La position 915 explore plus de types lorsque le squelette du QM est utilisé. Cela est probablement dû à la taille accrue de la poche S_0 dans ce squelette. Enfin, la position 920 est assez contrainte car très enfouie, ce qui limite les types explorés. L'analyse de la co-

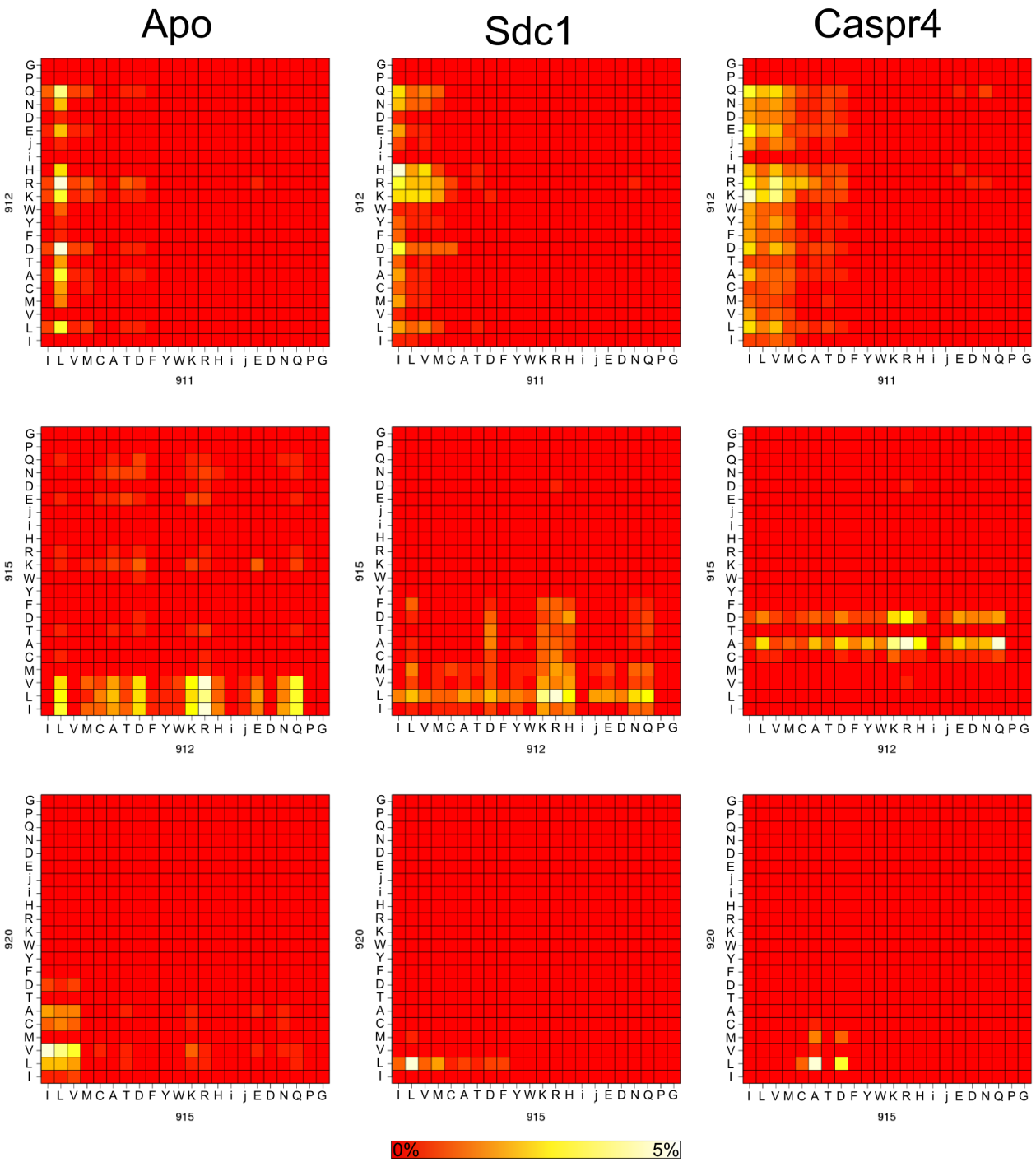


Figure 3.14 – Covariance des paires de positions lors des simulations de Monte Carlo à partir de squelette de la séquence native (4GVD). Les populations de chaque paire sont représentées par le dégradé de couleur, avec en jaune les paires les plus représentées ($\geq 5\%$) et rouge les moins représentées (0%).

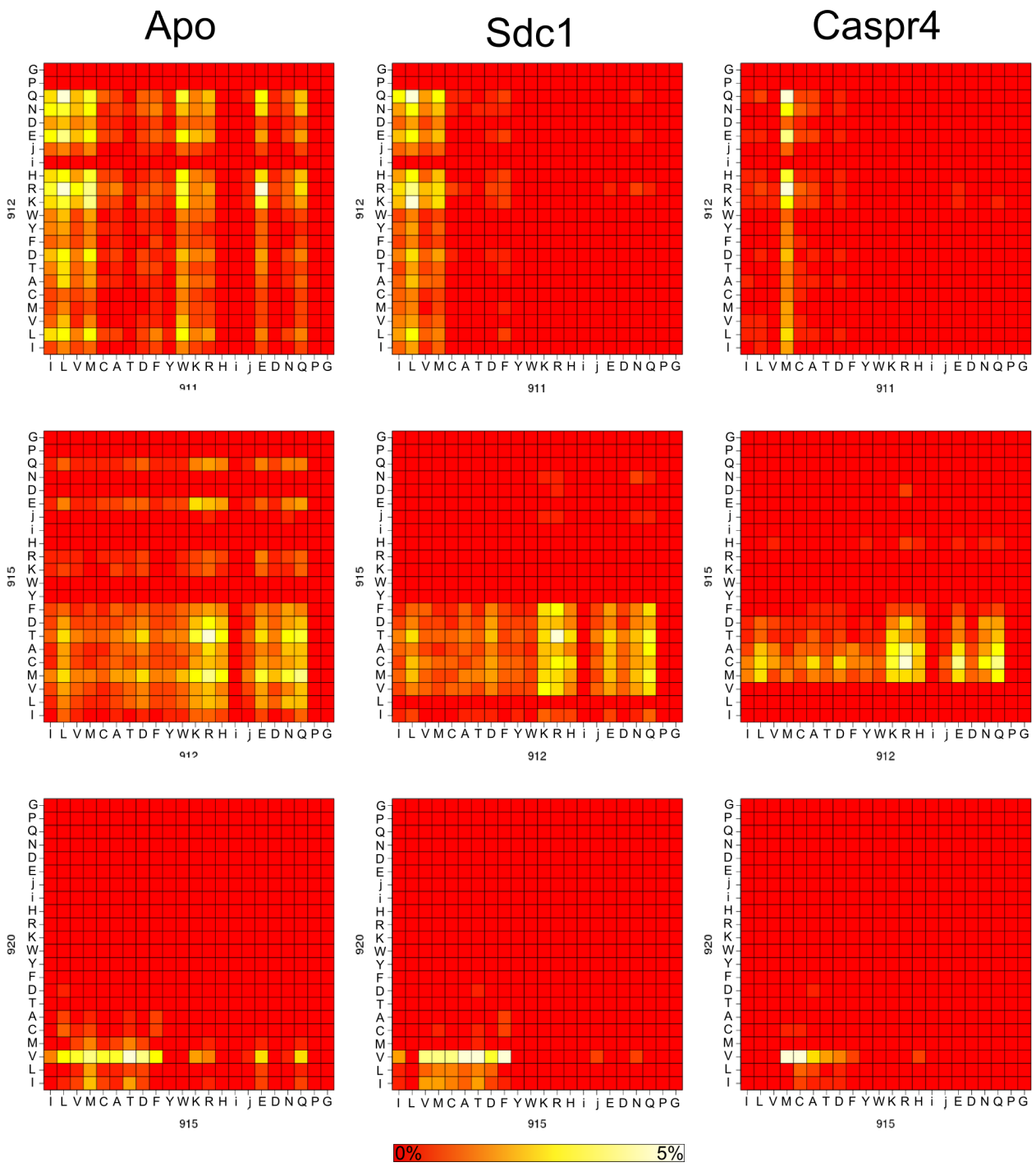


Figure 3.15 – Covariance des paires de positions lors des simulations de Monte Carlo à partir de squelette du quadruple mutant (4NXQ). Les populations de chaque paire sont représentées par le dégradé de couleur, avec en jaune les paires les plus représentées ($\geq 5\%$) et rouge les moins représentées (0%).

variance des positions montre principalement des lignes horizontales et verticales sans motifs diagonaux. Il ne semble donc pas y avoir de couplage fort entre les quatre positions mutées ce qui est cohérent avec les données expérimentales.

3.6 Discussion et conclusions

Dans ce chapitre nous avons optimisé les paramètres du modèle déplié pour redessiner les séquences des domaines PDZ. Les énergies de référence ont été ajustées pour reproduire les compositions expérimentales de deux domaines PDZ, Tiam1 et Cask, et de leurs homologues proches. L'approche utilisée présente deux nouveautés, en séparant les positions en deux groupes distincts (exposé/enfoui), et en effectuant l'optimisation des énergies de référence par groupes et non plus par types.

L'utilisation de ces paramètres pour le dessin complet des domaines PDZ de Tiam1 et Cask donne des performances similaires au programme Rosetta. Malgré ces bons résultats, le modèle présente encore quelques limites. Certaines ne sont pas intrinsèques au programme Proteus et concernent la plupart des programmes de CPD. La première provient de l'utilisation de la stabilité comme critère de sélection. En effet, bien que cette information soit importante, elle n'est pas la seule à prendre en compte lors de la modification de séquences. D'autres informations comme la spécificité du repliement (Mach & Koehl [2013]), la protection contre l'agrégation ou la prise en compte d'informations fonctionnelles sont également importantes. Il est néanmoins intéressant de noter que parmi les séquences produites par Proteus et Rosetta, aucune n'a été classée dans une autre superfamille par *Superfamily*. Cela suggère que les séquences produites sont spécifiques au repliement des domaines PDZ. Il est également possible de prendre en compte des informations fonctionnelles, telle que la liaison de ligand, en conservant les types natifs aux positions impliquées dans l'interface de liaison par exemple.

Une seconde limitation du modèle provient de l'utilisation d'un squelette fixe. Bien que les atomes du squelette ne soient pas autorisés à bouger lors de l'exploration Monte Carlo, la flexibilité du squelette est traitée de manière implicite par l'utilisation d'une constante diélectrique pour le soluté supérieure à 1. Les valeurs de 4 et 8 permettent en effet de modéliser les réarrangements du squelette et des chaines latérales en réponse à une mutation ou un changement de rotamère. Malgré cette flexibilité implicite, le modèle ne permet pas au squelette de modifier sa structure tridimensionnelle suite à une répulsion stérique par exemple. La rigidité

du squelette pourrait en partie expliquer la faible diversité observée dans le cœur hydrophobe des séquences produites. Cette diversité est augmentée lorsque les séquences obtenues sur les squelettes de Tiam1 et Cask sont regroupées. En effet, la validation croisée a montré que le jeu de paramètres était sensible au squelette utilisé, et ce malgré la proximité structurale entre Tiam1, Cask, DLG2 et la Synténine.

Enfin, bien que les fréquences des différents groupes soient respectées, la composition des différents types ne l'est pas parfaitement et certains acides aminés sont sur-représentés quand d'autres sont sous-représentés. Ce déséquilibre pourrait nuire à la qualité des séquences produites en rendant certains types trop rares. Ces disparités nous ont déjà amené à scinder certains groupes pour rééquilibrer les fréquences. Une autre possibilité serait de retirer les contraintes au sein d'un groupe dans les derniers pas de l'optimisation ou d'effectuer l'optimisation sur un plus petit nombre de positions, qui seraient donc plus contraintes géométriquement.

L'application du modèle pour redessiner quatre positions impliquées dans la spécificité de Tiam1 et Tiam2 a donné des résultats encourageants. En effet, les simulations ont produits des séquences cohérentes avec les séquences naturelles et notamment la séquence sauvage ou des homologues proches. La qualité des prédictions reste toutefois assez sensible au squelette utilisé puisque ce dernier a tendance à introduire un biais vers sa séquence native. L'utilisation d'un squelette fixe limite également les séquences explorées. L'introduction de la flexibilité au cours de l'exploration permettrait donc probablement d'améliorer la qualité des résultats.

Les simulations ont toutefois permis d'estimer de manière qualitative les changements d'affinité liés aux différentes mutations et ce en très bon accord avec les données expérimentales. La seule limitation de cette approche est la nécessité d'échantillonner les séquences (ou des séquences homologues) dans les états apo et holo. Lorsque ces données sont disponibles, cette approche peut permettre d'estimer de manière simple et à grande échelle l'affinité de complexes protéine:ligand. Elle peut donc constituer une première étape dans l'identification de mutations d'intérêt pour améliorer l'affinité ou modifier la spécificité de la protéine cible.

Chapitre 4

Étude de la stabilité des séquences générées par Proteus à partir du squelette du domaine PDZ de Tiam1

Au chapitre précédent nous avons optimisé les énergies de référence du programme Proteus pour modéliser des séquences compatibles avec le repliement des domaines PDZ. Les paramètres optimisés nous ont permis de produire des séquences similaires à celles des domaines PDZ retrouvés dans le base de donnée Pfam. Bien que ces résultats soient encourageants, le modèle présente plusieurs limites dont l'utilisation d'un squelette fixe et d'un solvant implicite. Une étape de validation supplémentaire consiste à tester la stabilité des séquences Proteus par simulations de dynamique moléculaire en solvant explicite. Ce test, bien plus strict que les précédents, permet de vérifier la compatibilité de la séquence avec le squelette de la protéine.

Dans ce chapitre nous décrirons les différentes étapes allant de la sélection des séquences Proteus à étudier jusqu'aux simulations de ces dernières afin d'analyser leur stabilité. Enfin, pour cinq des séquences, des tests *in vitro* ont été effectués par l'équipe de E. Fuentes (Université de l'Iowa, USA) pour confirmer les résultats des simulations et proposer des modifications pouvant améliorer la stabilité de nos séquences. Nous verrons qu'une des séquences Proteus et trois de ses variants se replient bien à environ 50°C.

4.1 Génération des séquences

La première étape consiste à générer des séquences et des structures compatibles avec le repliement du domaine PDZ de Tiam1 en utilisant les jeux de paramètres optimisés.

4.1.1 Conservation des résidus de l'interface protéine-peptide

Le but final de cette étude est de synthétiser puis de tester *in vitro* le repliement des séquences sélectionnées pour valider notre modèle. En prévision de ces tests, nous avons fait le choix de contraindre 11 positions de l'interface protéine-peptide (figure 4.1). L'avantage de conserver cette interface est double. Cela permettra d'abord de tester indirectement le repliement du domaine par des mesures d'affinité. En effet, si la protéine est capable de lier le peptide Sdc1, il y a de fortes chances qu'elle soit repliée correctement. D'autre part, les études par résonance magnétique nucléaire (RMN) de Tiam1 en solution ont montré que la liaison du peptide permettait de stabiliser le domaine. Dans le cas où la séquence se replie mais serait peu stable, l'ajout d'un peptide pourrait augmenter sa stabilité et donc permettre d'effectuer les analyses expérimentales.

Le choix des résidus à contraindre a été effectué en se basant sur les positions conservées dans l'alignement Pfam, les données bibliographiques et la structure du complexe Tiam1:Sdc1. La position P_0 du peptide est l'une des plus importantes. Elle est reconnue au niveau d'une poche, appelée S_0, fortement hydrophobe dans le cas de Tiam1, formée par les résidus Y858, F860, L915 et L920. Il semble donc important de conserver ces résidus. K850 est également important car il interagit avec l'extrémité C-terminale du peptide par l'intermédiaire d'une molécule d'eau (Shepherd *et al.* [2010] ; Pedersen *et al.* [2016]). T857 forme une liaison hydrogène avec le groupement carboxylate du peptide. Au niveau de la position P_{-1}, Shepherd *et al.* [2011] ont montré que S861 interagissait avec la chaine latérale de Y1. La poche S_{-2}, formée par les résidus L911 et K912 est impliquée dans la reconnaissance de la position P_{-2} et doit donc être conservée. Enfin, les derniers résidus sélectionnés sont S908 et N876 qui interagissent respectivement avec les positions P_{-4} et P_{-6}. S908 fait une liaison hydrogène avec E4 en cours de la simulation et N876 est très conservée dans l'alignement Pfam et forme une liaison hydrogène avec K6 (Liu *et al.* [2013]).

4.1.2 Jeux d'énergies de référence utilisés

La valeur de la constante diélectrique du soluté, ϵ_P, modifie sensiblement la composition des résidus du cœur. Nous avons donc décidé de générer des séquences à partir des deux jeux d'énergies de référence optimisés précédemment, avec ϵ_P=4 et ϵ_P=8 (tableau 3.2).

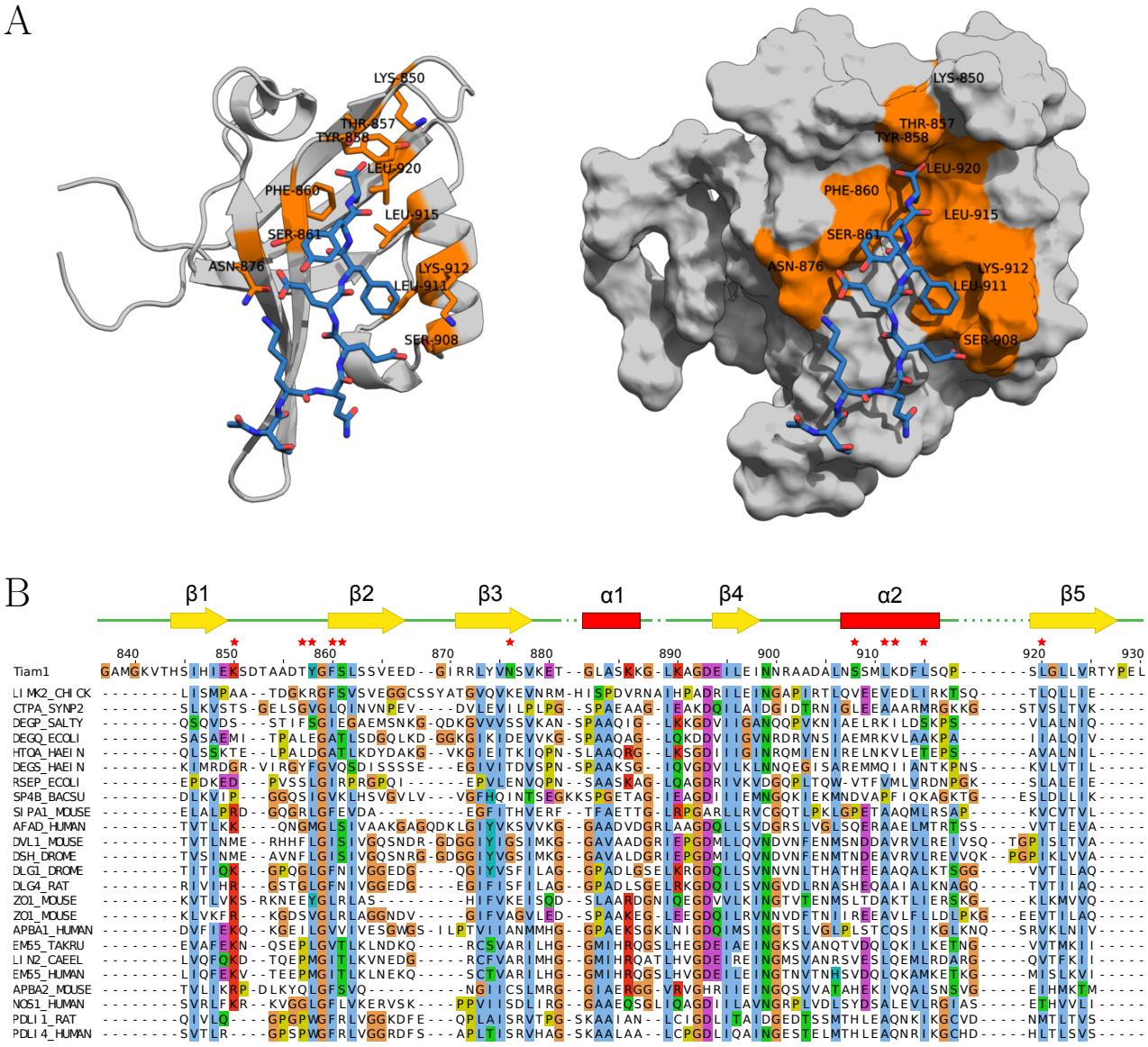


Figure 4.1 – Localisation des résidus de l'interface conservés lors de la génération des séquences. A : Les résidus fixés sont représentés en orange. Le peptide Sdc1 est représenté en bleu. B : La séquence de Tiam1 a été alignée sur quelques séquences de l'alignement Pfam. La position des résidus fixés est indiquée par les étoiles rouges.

4.2 Sélection des séquences

4.2.1 Critères de sélection des séquences à simuler

Les milliers de séquences générées par Proteus ne peuvent pas toutes être testées par dynamique moléculaire. Afin de sélectionner des séquences ayant de grandes chances d'être stables, nous appliquons des filtres pour extraire une dizaine de séquences. Ces filtres vérifient que les séquences sont proches des séquences PDZ naturelles et qu'on peut les produire dans

un organisme. On se base donc sur la comparaison avec les séquences expérimentales et sur les propriétés physico-chimiques des séquences Proteus.

Énergie Proteus Le premier filtre utilisé est l'énergie Proteus. Cette énergie correspond à l'énergie de dépliement de la protéine. Les 20000 meilleures énergies sont sélectionnées, ce qui représente une gamme d'énergie de 2 kcal/mol environ.

Score *Superfamily* Les résultats du programme *Superfamily* sont de bons indicateurs de la qualité des séquences générées. La sélection se fait sur trois critères. D'abord, elles doivent être reconnues comme appartenant à la famille des domaines PDZ. Ensuite, on les sélectionne sur la *E-value Superfamily*. Enfin, les séquences sont sélectionnées sur la longueur de la séquence reconnue par *Superfamily*.

Similarité et identité de séquences Nous supposons que les séquences produites ont d'autant plus de chances d'être stables qu'elles sont proches des séquences PDZ expérimentales. Les séquences sont donc comparées à la séquence native, aux homologues proches, et aux séquences provenant de l'alignement Pfam RP55 des domaines PDZ (12255 séquences). Pour les calculs de similarité, le matrice BLOSUM40 a été utilisée.

Nombre de mutations radicales La mutation d'un type vers un autre type présentant des propriétés physico-chimiques très différentes (taille, charge, etc...) peut déstabiliser la protéine. Nous définissons ainsi la notion de mutation radicale comme une mutation entre deux types d'acides aminés a et b possédant un score dans la matrice BLOSUM62 inférieure ou égale à -2.

Charge nette globale La présence d'un trop grand nombre de résidus chargés au sein de la protéine pourrait la rendre instable en raison de forces de répulsion importantes entre les groupements de même signe. Nous excluons donc les séquences dont la charge nette $C = \sum_{i=1}^{n} c_i \geq 6$, avec c_i la charge de l'acide aminé i à pH=7.

Point isoélectrique Le point isoélectrique (pI) d'une protéine représente le pH auquel la protéine est électriquement neutre. La valeur de pI dépend de la composition en acides aminés (à travers leurs propriétés acide/base), de la structure tridimensionnelle et de la nature du

solvant. Les séquences présentant un pI trop proche de 7 sont à exclure car, du fait de leur neutralité à pH physiologique, elles ont plus de chance de s'agréger *in vivo* si l'on souhaite les exprimer dans un organisme. La valeur du pI peut être approximée de manière théorique en se basant uniquement sur la composition en acides aminés :

$$pI \approx \frac{1}{n} \sum_{i=1}^{n} pK_{a,i} \tag{4.1}$$

où $pK_{a,i}$ correspond au pK_a théorique du résidu i et n au nombre de résidus titrables dans la séquence. En pratique, on utilise la valeur usuelle du pK_a des résidus titrables. Une autre solution, plus rigoureuse, consiste à reconstruire les modèles structuraux des différentes séquences puis à calculer le pI à l'aide du logiciel PROPKA (Olsson *et al.* [2011]). Cette méthode étant beaucoup plus couteuse en temps de calcul, elle n'a été appliquée que lorsque les valeurs des séquences sélectionnées étaient proches de 7.

4.2.2 Sélection automatique

L'utilisation de l'énergie Proteus comme premier filtre permet d'obtenir respectivement 2508 et 2712 séquences différentes pour les modèles $\epsilon_P = 4$ et $\epsilon_P = 8$. Les séquences ainsi retenues sont ensuite analysées de manière automatique à l'aide des critères ci-dessus. Pour chacun de ces critères il faut choisir une valeur seuil. Les seuils pour *Superfamily* et les mutations radicales sont fixés au premier quartile, ceux des scores de similarité au troisième quartile et les séquences ayant un pI $= 7 \pm 1{,}5$ sont exclues. Cela permet de conserver un petit nombre de séquences pour chaque filtre. Les résultats obtenus sont présentés dans le tableau 4.1. À l'issue de cette sélection, 45 et 66 séquences ($\epsilon_P = 4$ et $\epsilon_P = 8$ respectivement) remplissent tous les critères, soit un nombre de séquences pouvant être analysé manuellement.

4.2.3 Sélection manuelle

Le nombre de séquences sélectionnées reste trop élevé pour pouvoir tester la stabilité de chaque séquence par de longues simulations de dynamique moléculaire. Une deuxième sélection est donc effectuée manuellement pour garder 10 séquences. Les critères utilisés sont la présence ou non de cavité au sein du cœur hydrophobe, la charge de la protéine, la diversité des séquences et la prédiction des structures secondaires.

Tableau 4.1 – Critères de sélection des séquences Proteus produites

Descripteur	Valeur seuil	Nombre de séquences
Modèle $\epsilon_P = 4$		
Énergie Proteus	$\geq$ -27,2 kcal/mol	2508
Superfamily	E-value : $\leq 7{,}1.10^{-6}$, taille : $\geq$54	537/2508
Mutations radicales	≤ 14	791/2508
pI	$\neq 7 \pm 1{,}5$	1534/2508
S_{Pfam} [a]	$\geq$ -62	599/2508
$S_{Hom.}$ [b]	≥ 174	671/2508
Total		**45/2508**
Modèle $\epsilon_P = 8$		
Énergie Proteus	$\geq$ -85,8 kcal/mol	2712
Superfamily	E-value : $\leq 2{,}6.10^{-5}$, taille : $\geq$62	205/2712
Mutations radicales	≤ 15	923/2712
pI	$\neq 7 \pm 1{,}5$	1003/2712
S_{Pfam} [a]	$\geq$ -62	669/2712
$S_{Hom.}$ [b]	≥ 175	702/2712
Total		**66/2712**

[a] : Score de similarité par rapport à Pfam RP55
[b] : Score de similarité par rapport aux homologues

Toutes les séquences produites présentent une charge plus importante que la séquence sauvage. Pour se rapprocher le plus possible de la charge de cette dernière tout en conservant une bonne variabilité au sein des séquences, seules celles ayant une charge de +3 et +4 pour les séquences $\epsilon_P = 4$ et +4 et +5 pour les séquences $\epsilon_P = 8$ sont conservées.

La fonction d'énergie utilisée par Proteus peut entrainer des défauts de *packing* dans le cœur hydrophobe qui peuvent nuire à la stabilité de la protéine. Une inspection visuelle a montré que les mutations I848C et L906C créent une cavité de la taille d'une molécule d'eau au sein de la protéine. Cette cavité peut toutefois être comblée dans le cas de la mutation I848C par la mutation L922M. Les séquences présentant les mutations simples I848C et L906C sont donc éliminées.

Afin de vérifier la compatibilité des séquences avec le pli des domaines PDZ, les structures secondaires de chaque séquence sont prédites à l'aide des programmes Porter 4.0 (Pollastri & McLysaght [2004]) et YASPIN (Lin *et al.* [2004]) puis comparées au profil DSSP de la structure de Tiam1 (figure B.1 en annexe). Les deux programmes identifient correctement le positionnement des éléments de structure secondaire sur les séquences sélectionnées. Le programme

Tableau 4.2 – Valeurs des descripteurs des séquences sélectionnées.

Séquence	ϵ_P	Énergie Proteus	Superfamily E-value	Superfamily taille	%ID$_{WT}$	S$_{homo}$	S$_{pfam}$	Mutations radicales	Charge	pI
Tiam1	-	-	$2{,}23\times10^{-11}$	73	100,0	492,2	-22,8	0	-2	5,6
1	4	-26,83	$2{,}52\times10^{-6}$	54	36,2	180,9	-61,3	14	3	8,7
1’	4	-	$1{,}98\times10^{-5}$	51	37,2	187,0	-61,4	14	3	8,7
2	4	-26,41	$9{,}50\times10^{-7}$	55	36,2	181,0	-56,2	13	4	9,0
2’	4	-	$7{,}70\times10^{-6}$	52	37,2	187,1	-56,3	13	4	9,0
3	4	-27,13	$1{,}52\times10^{-6}$	55	38,3	182,4	-57,6	14	4	9,0
4	8	85,93	$2{,}20\times10^{-7}$	70	37,2	187,0	-52,1	15	4	8,8
4’	8	-	$1{,}66\times10^{-6}$	67	38,3	193,1	-52,2	15	4	8,8
5	8	86,18	$1{,}52\times10^{-5}$	72	39,4	201,5	-56,9	12	5	8,9
5’	8	-	$2{,}02\times10^{-4}$	71	40,4	207,6	-57,0	12	5	8,9
6	8	85,95	$1{,}83\times10^{-6}$	64	36,2	178,8	-58,3	13	5	8,9

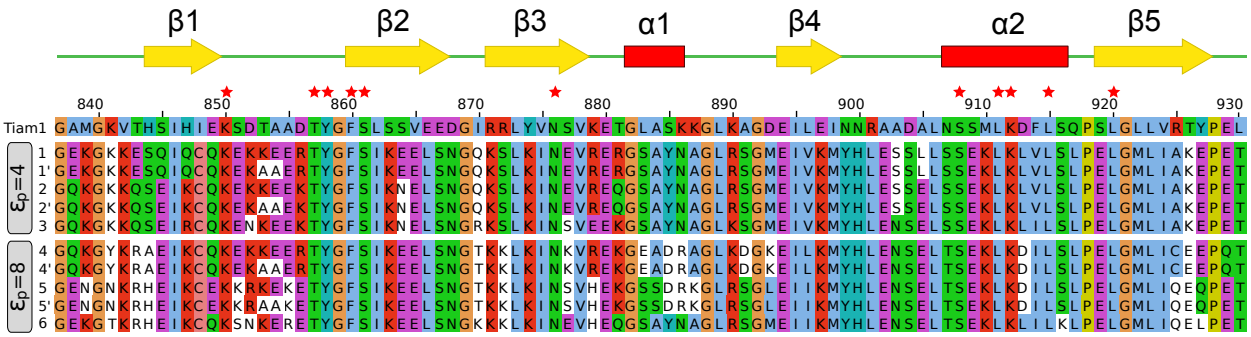


Figure 4.2 – Alignement des séquences sélectionnées. Les positions des résidus fixés sont indiquées par les étoiles rouges.

YASPIN semble cependant surestimer la longueur des hélices. Les séquences Proteus semblent donc compatibles avec la structure secondaire de Tiam1.

À l’issue de ces étapes, six séquences représentatives des deux protocoles ($\epsilon_P = 4$ et $\epsilon_P = 8$) et présentant des charges différentes sont sélectionnées. Pour quatre d’entre elles, la boucle 852-856 n’est composée que de résidus chargés, ce qui pourrait nuire à la stabilité de la protéine. Ces quatre séquences sont donc modifiées manuellement pour se rapprocher de la séquence sauvage (*ie* KKEEK → KAAEK). Les séquences modifiées seront par la suite indiquées par le symbole «’». Cela amène à 10 le nombre de séquences à tester par simulation de dynamique moléculaire. Les séquences sélectionnées ainsi que les descripteurs associés sont présentés en figure 4.2 et dans le tableau 4.2

4.3 Stabilité des séquences produites en simulation de dynamique moléculaire

À partir des séquences Proteus générées, les modèles tridimensionnels sont reconstruits. La stabilité de ces structures est ensuite analysée par dynamique moléculaire. La séquence native ainsi que le quadruple mutant (QM), moins stable expérimentalement, sont utilisés comme références.

4.3.1 Préparation des systèmes

Les structures générées sont solvatées dans une boîte d'eau explicite TIP3P octaédrique avec une distance minimale entre la protéine et les bords de la boîte de 13 Å, puis la charge du système est neutralisée à l'aide d'ions Na^+ ou Cl^-. Le système est ensuite minimisé pendant 1000 pas en imposant des contraintes harmoniques sur le squelette et en relâchant progressivement les contraintes sur les chaines latérales. On utilise successivement les algorithmes de *Steepest descent* et Newton-Raphson. Le champ de force AMBER ff99SB est utilisé, les liaisons covalentes impliquant des hydrogènes sont maintenues fixes à l'aide de l'algorithme SHAKE (Ryckaert *et al.* [1977]). Le pas d'intégration est de 2 fs. Les simulations sont maintenues à température et pression ambiantes par un thermostat et un barostat de Nose-Hoover (Nosé [1984] ; Hoover [1985]). Les interactions électrostatiques à longue portée sont traitées par la méthode *Particule Mesh Ewald* (PME, Cornell *et al.* [1996]). Le système est équilibré pendant 200 ps en augmentant progressivement le pas d'intégration et la température jusqu'aux valeurs cibles, tout en relâchant progressivement les contraintes sur le système. Une fois les systèmes équilibrés, des dynamiques de 100 ns sont effectuées à l'aide du programme NAMD 2.12 (Phillips *et al.* [2005]). Les simulations les plus stables sont prolongées, pour des durées totales comprises entre 200 ns et 1200 ns.

4.3.2 Critères de stabilité

Pour analyser la stabilité des séquences, plusieurs descripteurs sont comparés aux valeurs observées pour les formes apo et holo de la séquence native et du quadruple mutant. Le quadruple mutant est expérimentalement moins stable que la séquence native puisqu'il présente une énergie libre de dépliement de 0,82 ± 0,09 kcal/mol contre 2,60 ± 0,13 kcal/mol pour la

séquence native (Liu *et al.* [2016]). Ainsi, à température ambiante une large fraction de ce mutant est déstructurée (environ 20%).

4.3.2.1 Convergence des simulations

La convergence des simulations est mesurée en calculant l'écart quadratique moyen (RMSD pour *Root Mean Square Deviation*). Cette mesure permet d'étudier les fluctuations du système autour d'une conformation de référence. Dans notre cas, nous utilisons deux structures de référence distinctes. La première correspond à la structure initiale qui est donc très proche de la structure cristallographique de Tiam1 et qui permet de voir si la structure au cours de la simulation s'éloigne de la structure native. Mais cet éloignement ne signifie pas forcément que la structure est instable. En effet, la modification de la séquence peut stabiliser une conformation différente de la conformation native. C'est pourquoi une deuxième mesure du RMSD est effectuée en prenant comme référence la structure la plus proche de la structure moyenne de la dynamique. Le RMSD est calculé sur les atomes lourds du squelette protéique. Les extrémités de la protéine (résidus 837-841 et 927-930) étant très flexibles, ils sont exclus du calcul pour ne pas bruiter les résultats. De même, une seconde mesure du RMSD est effectuée en excluant la boucle flexible β_1-β_2 (résidus 851-856).

4.3.2.2 Fluctuations atomiques

Le RMSF (*Root Mean Square Fluctuation*) autour de la structure moyenne décrit les fluctuations atomiques au cours de la simulation pour chaque atome du système. Nous nous intéresserons à l'atome Cα de chaque résidu. Cette mesure permet de localiser des régions de la protéine dont la dynamique serait modifiée dans les séquences Proteus.

4.3.2.3 Rayon de giration

Le rayon de giration, noté Rg, donne une mesure de la compacité du système et peut être un indicateur des changements structuraux et du dépliement partiel ou complet de la protéine. Il s'écrit :

$$R_g = \sqrt{\frac{\sum_i m_i (r_i - r_{CM})^2}{\sum_i m_i}} \tag{4.2}$$

où r_{CM} est la position du centre de masse de la protéine, r_i est la position et m_i la masse de l'atome i.

4.3.2.4 Stabilité des structures secondaires

Le repliement de la protéine dépend en grande partie de la présence des structures secondaires. De ce fait, ces dernières constituent un très bon indicateur de stabilité. Pour suivre la stabilité des structures secondaires au cours des simulations, le programme DSSP est utilisé (Kabsch & Sander [1983]). Ce dernier se base sur la présence de liaisons hydrogène pour déterminer si un résidu est impliqué dans une structure secondaire ou non.

4.3.2.5 Analyse en composantes principales (ACP)

Afin de dégager les mouvements dynamiques globaux, l'analyse en composantes principales (ACP) est utilisée. Cette approche réduit la dimension des données en identifiant des combinaisons linéaires des coordonnées, appelées composantes principales (PC), qui préservent la variance des données. Elle permet de mettre en évidence l'existence de mouvements globaux et concertés.

La matrice de covariance des coordonnées des atomes Cα est calculée et pondérée par les masses atomiques, puis diagonalisée. Les valeurs propres correspondent à des déplacements corrélés et les vecteurs propres à leur amplitude. À partir de ces composantes, plusieurs analyses peuvent être effectuées. Il est notamment possible de projeter les conformations extraites de la trajectoire dans un plan défini par deux composantes pour étudier la réversibilité des mouvements de la protéine. La contribution de chaque résidu peut également être calculée pour déterminer quels résidus sont les plus flexibles. Enfin, la direction et l'amplitude des mouvements de chaque composante peuvent être visualisées sur la structure tridimensionnelle de la protéine.

4.4 Comportement des simulations de dynamique moléculaire

4.4.1 Convergence et stabilité des simulations

Les valeurs des RMSD au cours de simulations ont tout d'abord été étudiées pour les séquences sauvage et QM de Tiam1. Cette étape permet d'identifier si les deux séquences se comportent de manière similaire ou si l'instabilité de mutant se répercute sur les valeurs du

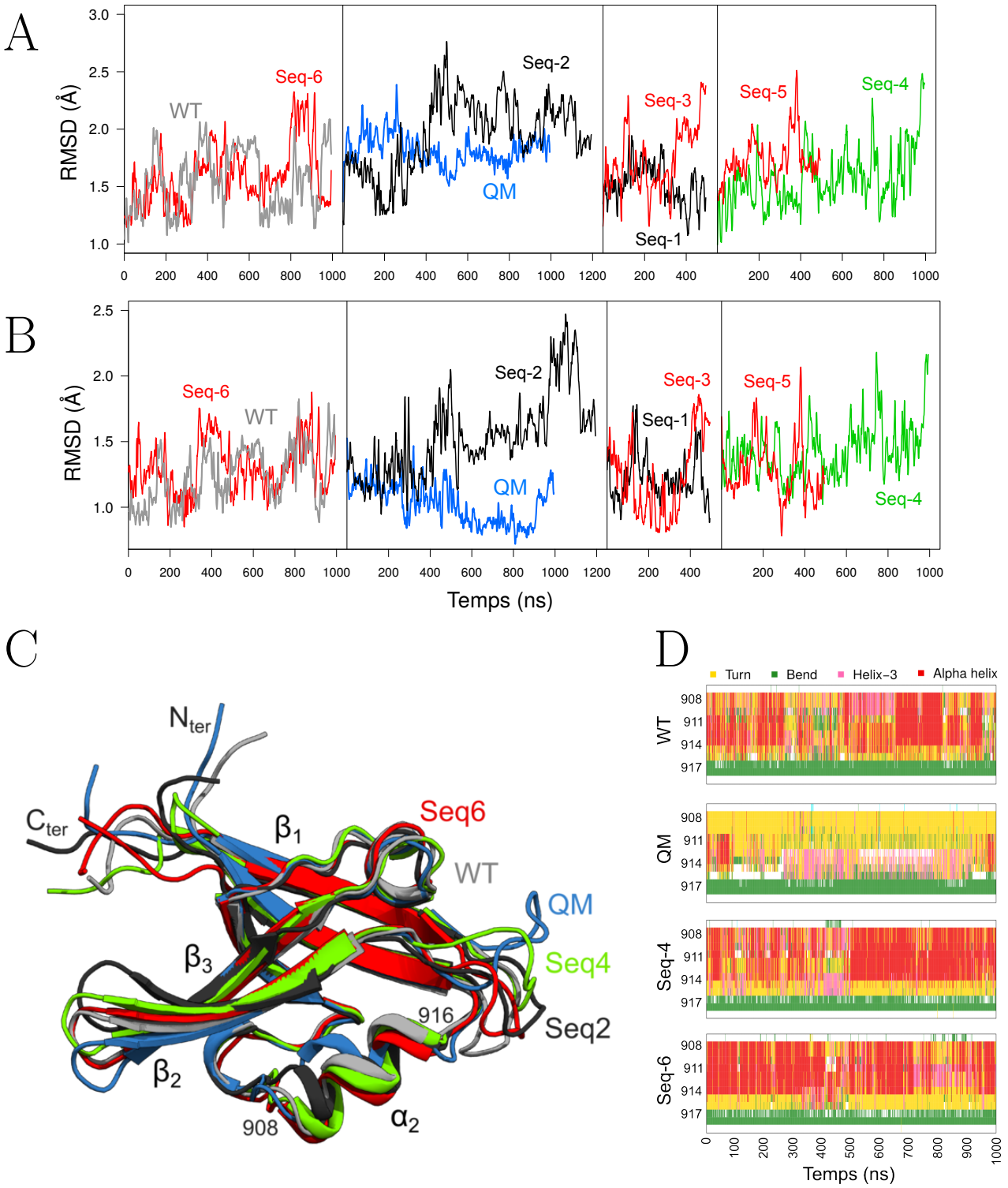


Figure 4.3 – Simulations des variants produits par Proteus. (A) RMSD des atomes du squelette des séquences WT et QM et de 6 variants par rapport à la structure initiale. (B) RMSD rapport à la structure moyenne. (C) Structures moyennes des séquences WT, QM, seq2, seq4 et seq6. (D) Structure secondaire calculée par DSSP au cours des simulations.

RMSD. La séquence sauvage est stable durant toute la simulation de 1 μs (figure 4.3A). Le RMSD de la structure moyenne par rapport à la structure cristallographique est de 1 Å lorsque l'on exclue les résidus des extrémités. Au cours de la trajectoire, le RMSD moyen par rapport à la structure moyenne varie entre 1 et 1,5 Å sans dérive apparente (figure 4.3B). La boucle β_1-β_2 modifie peu les valeurs de RMSD ce qui indique qu'elle est stable. Le profil DSSP indique que les structures secondaires sont conservées tout au long de la simulation avec néanmoins quelques instabilités au niveau de l'hélice α_2.

Le profil de la séquence QM est assez différent du sauvage. En effet, la structure moyenne du QM présente un RMSD par rapport à la structure cristallographique de 1,6 Å. Cette valeur plus élevée s'explique principalement par l'instabilité de l'hélice α_2 qui disparait dans le profil DSSP (figure 4.3D). Le RMSD par rapport à la structure moyenne est stable et compris entre 0,8-1,2Å. Au cours de la simulation, la structure du QM s'écarte donc de la structure initiale pour conserver par la suite une conformation dans laquelle l'hélice α_2 est déstructurée (figure 4.3C).

Les dix séquences Proteus conservent leurs structures secondaires tout au long des simulations. De même, les valeurs des rayons de giration sont comparables à celles des WT et QM. Les structures moyennes sont toutes très semblables à la structure moyenne de WT (figure 4.3C). Malgré cette stabilité apparente, l'analyse des RMSD révèle des profils très différents de la séquence sauvage. En effet, à part la séquence 4, toutes les séquences atteignent rapidement des valeurs de RMSD par rapport à la structure cristallographique entre 1,9 Å et 2,4 Å contre 1,6 Å pour WT. Cette augmentation est principalement due à la boucle β_1-β_2 puisque son exclusion lors des calculs de RMSD permet de retrouver des valeurs proches de celle de WT, 1,5 Å et 2,0 Å. Cependant les RMSD des séquences 2, 3 et 4' augmentent au cours des simulations. Les séquences 2 et 3 sont par ailleurs les seules séquences dont le RMSD par rapport à la structure moyenne n'est pas stable. Quant à la séquence 4', sa structure moyenne possède un RMSD par rapport à la structure cristallographique élevé principalement en raison du déplacement de la boucle β_2-β_3 et d'une rotation de l'hélice α_2.

Parmi les trois séquences simulées jusqu'à 1 μs (les séquences 2, 4 et 6), la séquence 2 apparait comme la moins stable et présente notamment un pic important du RMSD par rapport à la structure moyenne entre 950 ns et 1100 ns. Cette augmentation brutale est causée par un déplacement de la région 875-895 et une déformation de α_2. La prolongation de la dynamique à 1200 ns permet néanmoins d'observer une diminution du RMSD. Il est donc

difficile de savoir si ces fluctuations sont le signe de l'instabilité de la séquence. La séquence 4 présente également des fluctuations plus importantes en fin de simulation. La séquence 6 semble être la plus stable des trois séquences. Une augmentation du RMSD est tout de même observée au cours de 200 dernières nanosecondes et est causée par le déplacement de la boucle β_1-β_2 qui entraine un déplacement de α_2. L'instabilité de l'hélice α_2 pourrait en partie provenir du champ de force ff99SB utilisé. En effet, plusieurs études ont montré que ce champ de force sous-estime la stabilité des hélices α (Best & Hummer [2009] ; Lindorff-Larsen *et al.* [2012] ; Cino *et al.* [2012] ; Kia & Darve [2013] ; Smith *et al.* [2015]).

4.4.2 Identification des résidus responsables des fluctuations

L'analyse des RMSD a mis en évidence trois régions dont la flexibilité semble accrue dans la plupart des séquences Proteus mais également dans le QM, à savoir les boucles β_1-β_2 et β_2-β_3 et l'hélice α_2. Les fluctuations de ces régions pourraient traduire une stabilité moindre des séquences, comme c'est le cas du QM. Il semble donc important d'identifier les résidus impliqués dans ces mouvements pour tenter, par la suite, de modifier manuellement les séquences afin d'augmenter leur stabilité.

L'instabilité de l'hélice α_2 dans les séquences Proteus est principalement due aux interactions des résidus E/N864 et S908/907 au cours des simulations qui courbent l'hélice (figure 4.4A). Dans la séquence sauvage, une sérine est présente à la position 864. Sa chaine latérale étant trop courte pour interagir avec ces sérines, l'hélice n'est pas déformée. Dans le cas de QM, la déstructuration de l'hélice α_2 est causée par des interactions entre les résidus E912/S908 d'une part et F860/F915 d'autre part (figure 4.4B). En effet, en interagissant avec S908, E912 modifie le pas de l'hélice entre les résidus 908 et 912. F915 interagit avec F860 par le biais d'une interaction de π-*stacking* qui déstructure l'hélice autour de la position 915 (figure 4.4C). Les positions 912 et 915 étant mutées dans le QM, leur effet sur la stabilité de α_2 pourrait en partie expliquer la faible stabilité de ce mutant *in vitro*.

Au cours des simulations des séquences 1, 1', 2, 2', 5 et 6, les résidus polaires de la boucle β_1-β_2 interagissent avec ceux des brins β_2 et β_3 (figure 4.4C). Cela semble confirmer que la présence de résidus chargés au niveau de cette boucle est défavorable à la stabilité de la protéine. De plus, en interagissant avec les brins β_2 et β_3, la boucle β_1-β_2 vient recouvrir une partie de l'interface de liaison du peptide. L'apparition de telles interactions dans les séquences

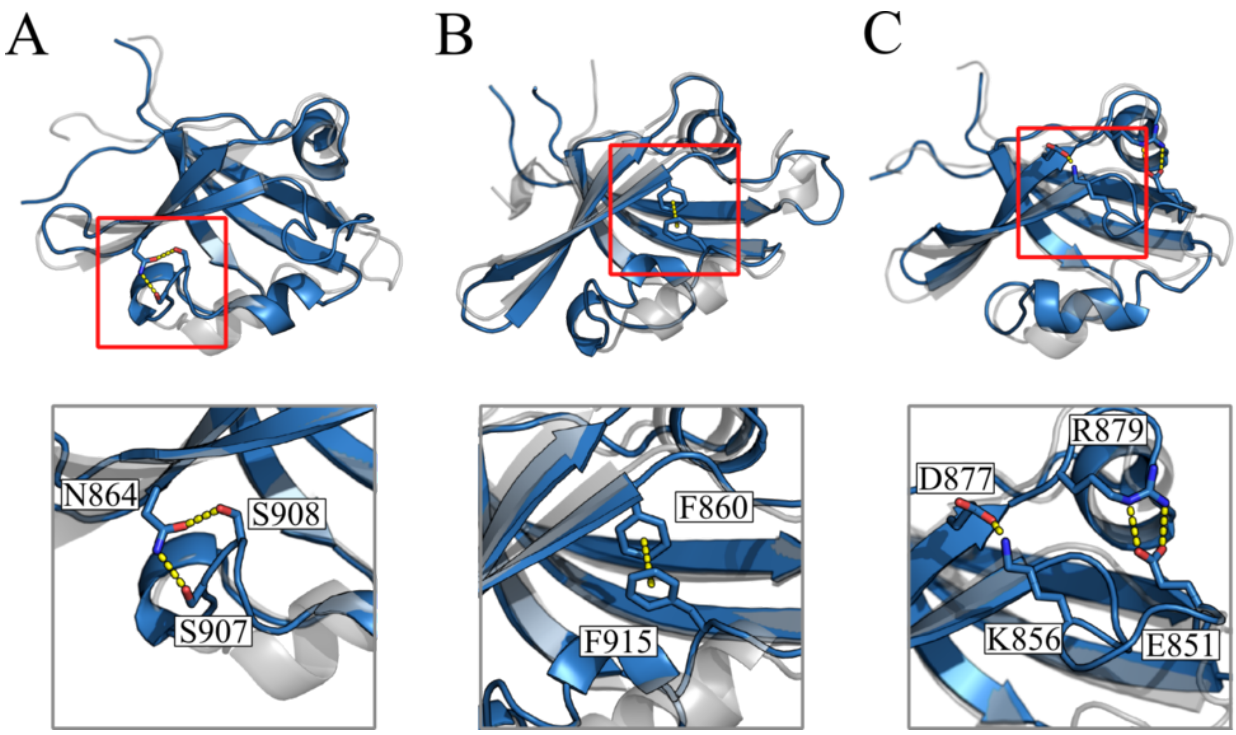


Figure 4.4 – Déformations des structures de Tiam1 QM et des séquences Proteus au cours des simulations de dynamique moléculaire. Les structures cristallographiques sont en gris. A : Déformation de α_2 au cours de la simulation de la séquence 2'. B : Déformation de l'hélice α_2 au cours de la simulation de QM. C : Déplacement de la boucle β_1-β_2 au cours de la simulation de la séquence 2'.

modifiées et non modifiées semble indiquer que les modifications manuelles de la boucle β_1-β_2 ne sont pas suffisantes.

4.4.3 Analyse en composantes principales

Les séquences 4 et 6 semblent être les plus stables et sont donc sélectionnées pour des analyses plus poussées. L'analyse en composantes principales identifie les mouvements collectifs de la protéine au cours de la dynamique. Cette analyse a été effectuée sur les atomes du squelette de la protéine Tiam1 sauvage et du quadruple mutant liés ou non à un peptide (Sdc1 et Caspr4 respectivement) ainsi que sur les séquences 4 et 6. Ces analyses permettront d'une part d'étudier l'impact du ligand sur la dynamique de Tiam1 mais également de savoir si les mouvements observés pour les séquences 4 et 6 sont plus proches de la séquences sauvage ou du quadruple mutant. Généralement, le nombre de composantes retenu est choisi afin de décrire 80% de la variance totale des positions atomiques au cours de la dynamique. Ici, nous

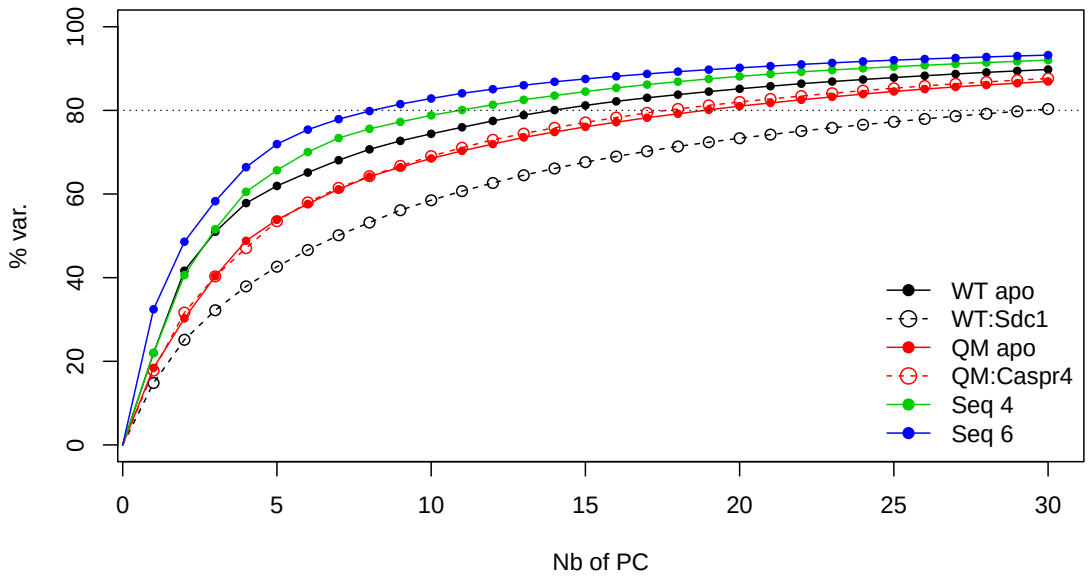


Figure 4.5 – Pourcentage de variabilité expliquée par les 30 premières composantes de l'ACP.. L'ACP a été effectuée sur les atomes lourds du squelette de chaque protéine.

avons fait le choix de nous limiter à 10 composantes, ce qui permet de décrire entre 69 et 83% de la variance (figure 4.5).

Afin d'analyser la convergence des simulations, les déplacements sont projetés sur les plans définis par les dix premières composantes prises deux à deux. Une trajectoire a convergé si le système fluctue réversiblement autour de la structure moyenne, ce qui se traduit par un nuage de points homogène balayé plusieurs fois. Dans la majorité des cas, nous observons ce comportement (figure 4.6). Seules les composantes 1 et 2 de WT apo et de la séquence 6 présentent deux ou trois groupes distincts, ou bassins, avec un petit nombre de transitions entre bassins. Dans ces deux cas, les simulations de 1 μs ne sont pas suffisantes pour bien échantillonner les transitions entre bassins.

Pour identifier et visualiser les mouvements décrits par les différentes composantes, la contribution de chaque résidu est calculée. L'amplitude des mouvements et leur direction sont également projetées sur la structure moyenne (figure 4.7). Comme attendu, les régions qui contribuent le plus aux composantes sont les trois régions les plus flexibles identifiées, à savoir les boucles β_1-β_2 et β_2-β_3, l'hélice α_2 et, dans une moindre mesure, la boucle β_3-α_1. Les premières composantes de WT apo sont caractérisées par des mouvements concertés de la boucle β_2-β_3 et de l'hélice α_2 (figure 4.7A). Ces mouvements ne sont pas observés dans

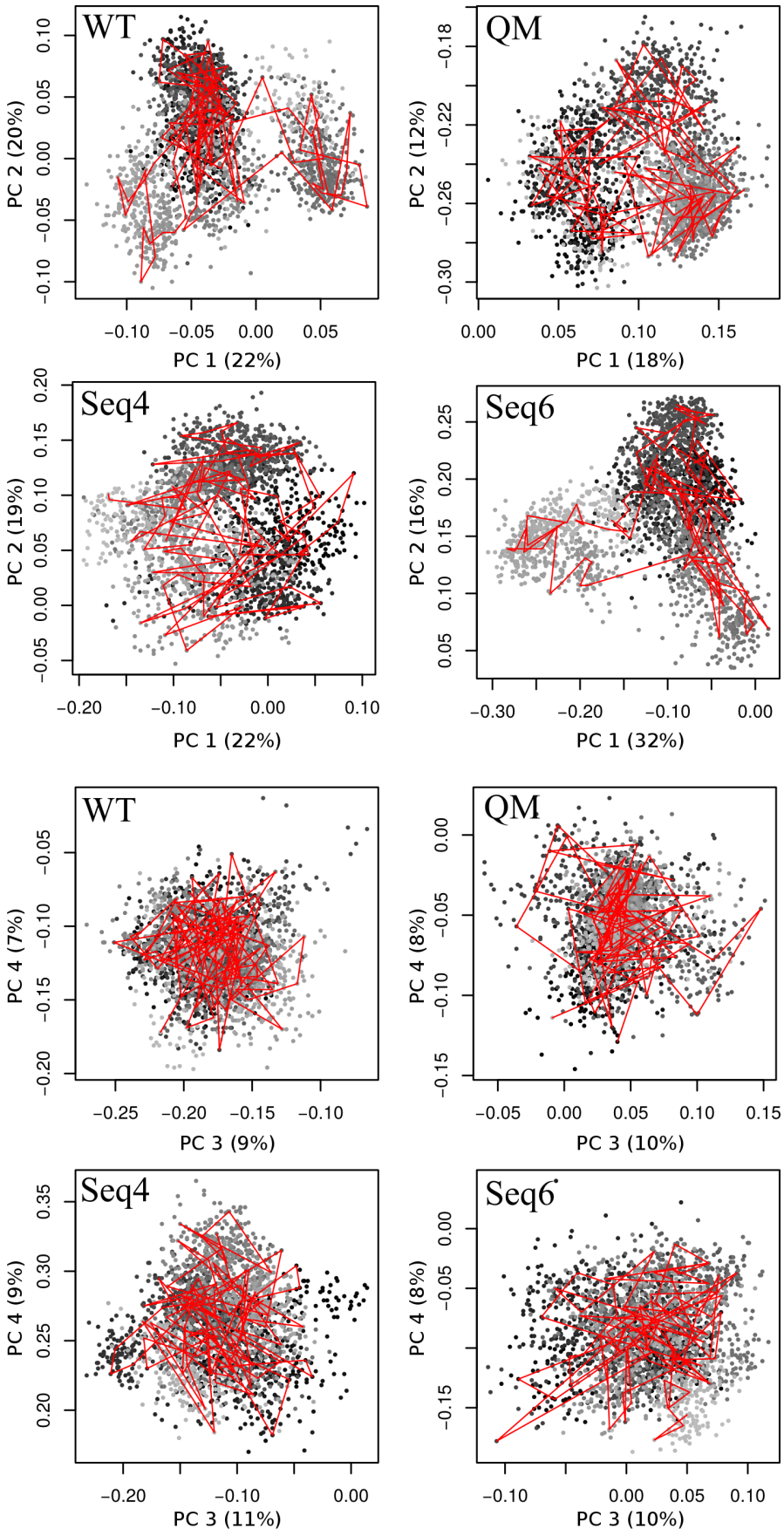


Figure 4.6 – Projections des trajectoires sur les composantes PC1 à PC4 de l'ACP. Chaque point représente une conformation. Points noirs (gris) : premières (dernières) 500 ns.

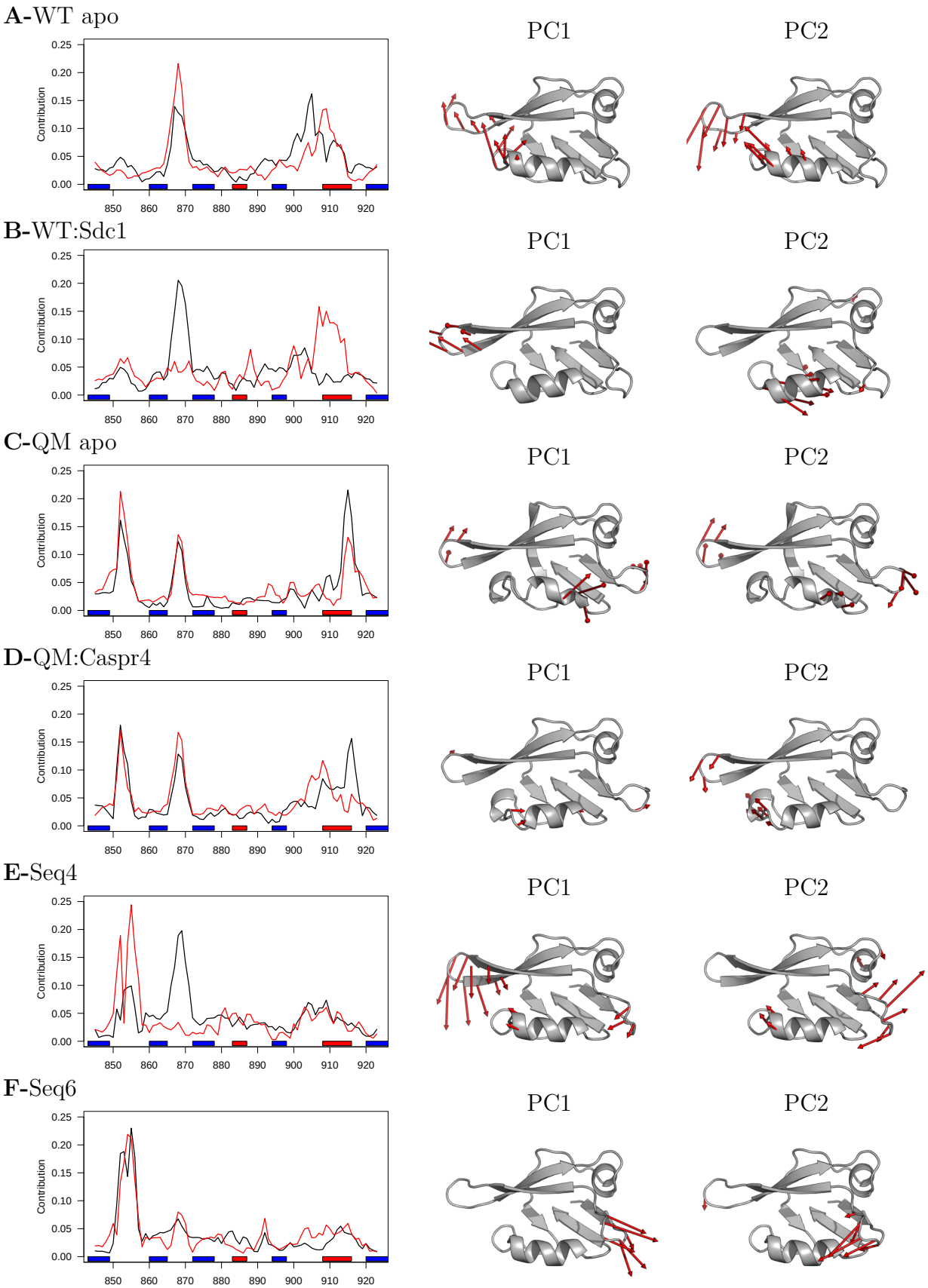


Figure 4.7 – Contribution des résidus aux deux premières composantes principales. Pour chaque séquence, la contribution des résidus aux composantes 1 (noire) et 2 (rouge) est représentée. La direction et l'amplitude des mouvements sont représentées par les flèches rouges.

le complexe WT:Sdc1 pour lequel la boucle β_2-β_3 et l'hélice α_2 présentent des mouvements décorrélés respectivement décrits par les composantes 1 et 2 (figure 4.7B).

Les formes apo et holo du quadruple mutant ont des mouvements similaires et notamment des mouvements corrélés des boucles β_1-β_2 et β_2-β_3 et de l'hélice α_2 dans les deux premières composantes (figure 4.7C et D). Pour les deux dynamiques du quadruple mutant, l'hélice α_2 présente une dynamique plus importante, principalement autour de la position 915 qui est l'une des positions mutées.

Pour les séquences 4 et 6, les régions présentant des fluctuations importantes sont les boucles β_1-β_2 et β_2-β_3 (figure 4.7E et F) comme chez WT. Contrairement à WT, les mouvements de ces boucles sont décorrélés de l'hélice α_2. Cette différence pourrait être due aux fluctuations importantes de la boucle β_2-β_3 qui masqueraient les mouvements de moins grande ampleur de α_2. La dynamique des séquences 4 et 6 semble donc plus proche de celle de WT que de QM. Cela peut refléter soit une meilleure stabilité de ces séquences par rapport à QM, soit une dépendance de la dynamique à la structure de départ, les séquences ayant été produites sur le squelette de WT.

4.4.4 Comparaison des dynamiques aux données RMN

Les séquences 4 et 6 seront étudiées *in vitro*, notamment par RMN. Il est donc intéressant d'extraire des simulations une mesure directement comparable. Pour ce faire, le paramètre d'ordre de Lipari-Szabo, S^2, a été calculé pour identifier les zones les plus flexibles. L'approche utilisée est celle du modèle libre (Koller *et al.* [2008] ; Bowman [2016]) qui relie la densité spectrale mesurée par RMN à la fonction d'autocorrélation d'un vecteur unitaire le long d'une liaison atomique :

$$S^2 = \lim_{t\to\infty} C(t) = \lim_{t\to\infty} \langle P_2(\hat{\mu}(0).\hat{\mu}(t))\rangle \tag{4.3}$$

où $\hat{\mu}$ est le vecteur normalisé formé par deux atomes au cours de la trajectoire et P_2 est un polynôme de Legendre d'ordre 2 :

$$P_2 = \frac{1}{2}(3x^2 - 1) \tag{4.4}$$

S^2 peut ainsi être calculé à partir de la dynamique de la manière suivante :

$$S^2 = \frac{2}{3}\Big[\langle\hat{\mu}_x^2\rangle^2 + \langle\hat{\mu}_y^2\rangle^2 + \langle\hat{\mu}_z^2\rangle^2 + \langle\hat{\mu}_x\hat{\mu}_y\rangle^2 + \langle\hat{\mu}_x\hat{\mu}_z\rangle^2 + \langle\hat{\mu}_y\hat{\mu}_z\rangle^2\Big] - \frac{1}{2} \tag{4.5}$$

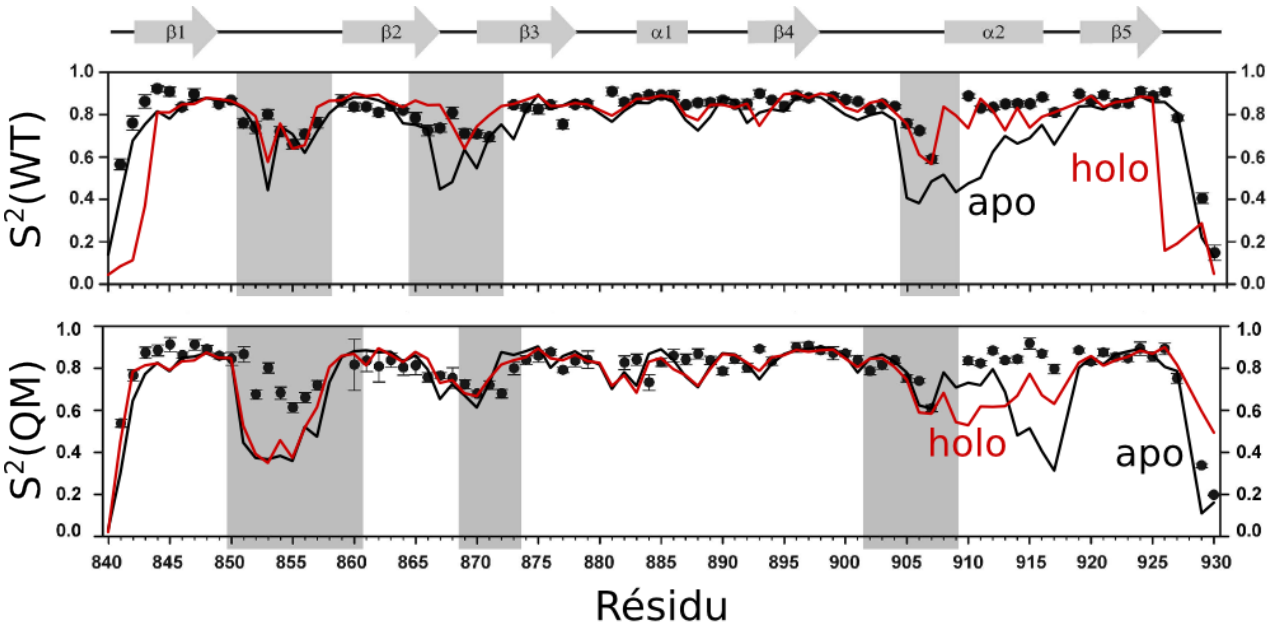


Figure 4.8 – Paramètres d'ordre S^2 obtenus par RMN (points) et dynamique moléculaire (traits) pour les NH du squelette en présence ou non d'un peptide (traits rouges et noirs).

avec $\hat{\mu}_x$ la composante x du vecteur unitaire $\hat{\mu}$.

La valeur S^2 varie entre 0 et 1. Une valeur proche de 0 signifie que la liaison ne reste pas dans la même orientation au cours de la simulation, ce qui correspond à une région flexible. Une valeur proche de 1 correspond à une liaison gardant la même orientation durant la dynamique et donc à un résidu très structuré. Les résidus impliqués dans des structures secondaires présentent généralement des paramètres d'ordre S^2 supérieurs à 0,85 pour les liaisons NH du squelette. Les régions non structurées ont des valeurs entre 0,4 et 0,6. Afin d'étudier la convergence des simulations, le terme S^2 est calculé sur la trajectoire complète et sur les deux moitiés de trajectoires prises séparément.

La méthode est testée sur les formes WT et QM de Tiam1, en présence ou non d'un peptide, systèmes pour lesquels des données expérimentales sont disponibles. Les valeurs des paramètres d'ordre obtenues sont en bon accord avec les données RMN (figure 4.8). Les trois régions les plus flexibles (α_2, β_1-β_2 et β_2-β_3) sont les mêmes mais leur flexibilité est surestimée dans les dynamiques. C'est particulièrement le cas pour l'hélice α_2, sauf dans la simulation de WT:Sdc1.

Pendant la simulation de WT apo, la boucle β_2-β_3 interagit avec l'extrémité N-terminale de la protéine. Cette interaction pourrait expliquer la flexibilité accrue de cette région. Les simulations des formes holo et apo du QM présentent également des fluctuations plus impor-

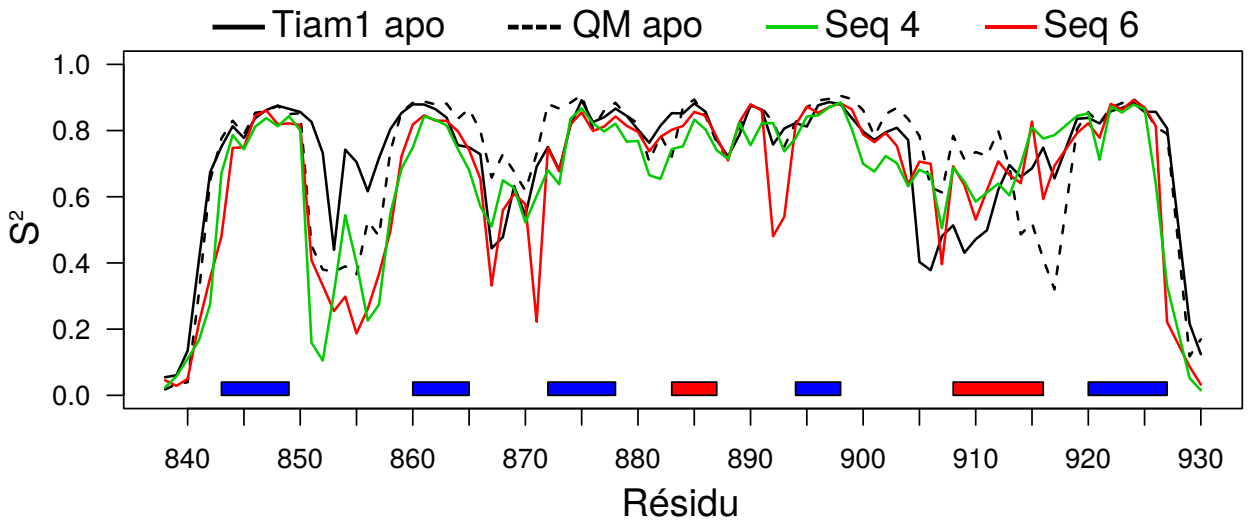


Figure 4.9 – Paramètres d'ordre S^2 obtenus des séquences 4 et 6.

tantes au niveau de l'hélice α_2. Le profil observé pour la boucle β_1-β_2 est très différent des données RMN avec une plus grande flexibilité. Il est à noter que la structure cristallographique utilisée pour les simulations du QM possède une boucle β_1-β_2 plus exposée au solvant. Cette conformation pourrait être responsable de la flexibilité accrue observée, la boucle étant plus libre. Enfin, la présence d'un peptide dans les formes holo du QM et WT semble stabiliser la protéine, principalement au niveau de l'hélice α_2, ce qui est également observé expérimentalement. L'analyse des deux moitiés de trajectoires séparément met en évidence de plus grandes disparités dans les formes apo que holo. Cela confirme le fait que les complexes sont plus stables. Dans WT apo, la boucle β_2-β_3 et l'hélice α_2 sont plus flexibles dans le seconde moitié de la simulation. Au contraire, β_1-β_2 est plus flexible dans la première moitié de la simulation du QM apo. Des valeurs systématiquement plus faibles dans la seconde moitié des trajectoires pourraient traduire une dépendance entre la stabilité et le temps de simulation, ce qui ne semble pas le cas ici. La flexibilité accrue de certaines régions dans les simulations suggère que, *a contrario*, une protéine stable en dynamique moléculaire a des chances d'être encore plus stable *in vitro*.

Les mêmes analyses sont appliquées aux séquences 4 et 6. Les profils observés sont comparables à ceux obtenus pour WT et QM apo (figure 4.9). Les valeurs de S^2 sont néanmoins plus faibles pour la boucle β_1-β_2, principalement à cause des mouvements causées par les résidus chargés décrits précédemment. L'hélice α_2 semble moins flexible que dans les formes apo du WT et du QM mais les valeurs observées restent inférieures aux valeurs expérimentales. Comme pour les séquences WT et QM, l'analyse des deux moitiés des trajectoires ne montre

pas une flexibilité accrue dans la seconde partie de la trajectoire. Au contraire, la séquence 4 semble plus stable autour des résidus 880 (boucle β_1-β_2) et 912 (hélice α_2) dans la seconde moitié de la simulation. Quant à la séquence 6, les principales différences se situent au niveau des boucles β_1-β_2 et α_1-β_4 qui sont plus flexibles dans la seconde et première moitié de la trajectoire respectivement.

4.5 Stabilité des séquences *in vitro*

Toutes les expériences ont été réalisées par Y.J. Sun du laboratoire de E. J. Fuentes (Université de l'Iowa, USA).

4.5.1 Analyse de la séquence 6 par RMN

La séquence 6 a été exprimée dans *E. coli* puis purifiée par chromatographie d'exclusion stérique (SEC). Les résultats montrent que contrairement à la séquence sauvage, seq6 a tendance à former des multimères et à précipiter. Pour déterminer si seq6 se replie, des expériences de dichroïsme circulaire (DC) et de RMN 1-D ont été effectuées. Ces deux méthodes permettent d'estimer si la protéine est structurée ou non. Dans le cas du DC, la composition en hélices α et feuillets β peut être évaluée à partir des propriétés optiques de la protéine. Les résultats RMN indiquent que la séquence se replie partiellement et présente des structures hélicoïdales (figure 4.10A). La présence de bandes à 222 et 208 nm dans le profil DC, caractéristiques des hélices α et à 218 nm caractéristique des feuillets β (figure 4.10B) confirment les résultats RMN. La protéine est cependant partiellement dépliée comme en témoignent les bandes autour de 195 nm. Le profil DC est tout de même assez éloigné du WT. L'effet des peptides Sdc1 et Caspr4 sur la stabilité de la protéine a également été étudié par fluorimétrie différentielle à balayage qui permet de déterminer le point de fusion de la protéine. Dans le cas où le peptide stabiliserait la protéine, sa présence devrait augmenter la valeur du point de fusion. La présence d'un peptide ne modifie pas le point de fusion de seq6 qui est compris entre 50,2 et 50,5°C (figure 4.10). Cela signifie que seq6 se replie et est stable mais est incapable de lier un peptide.

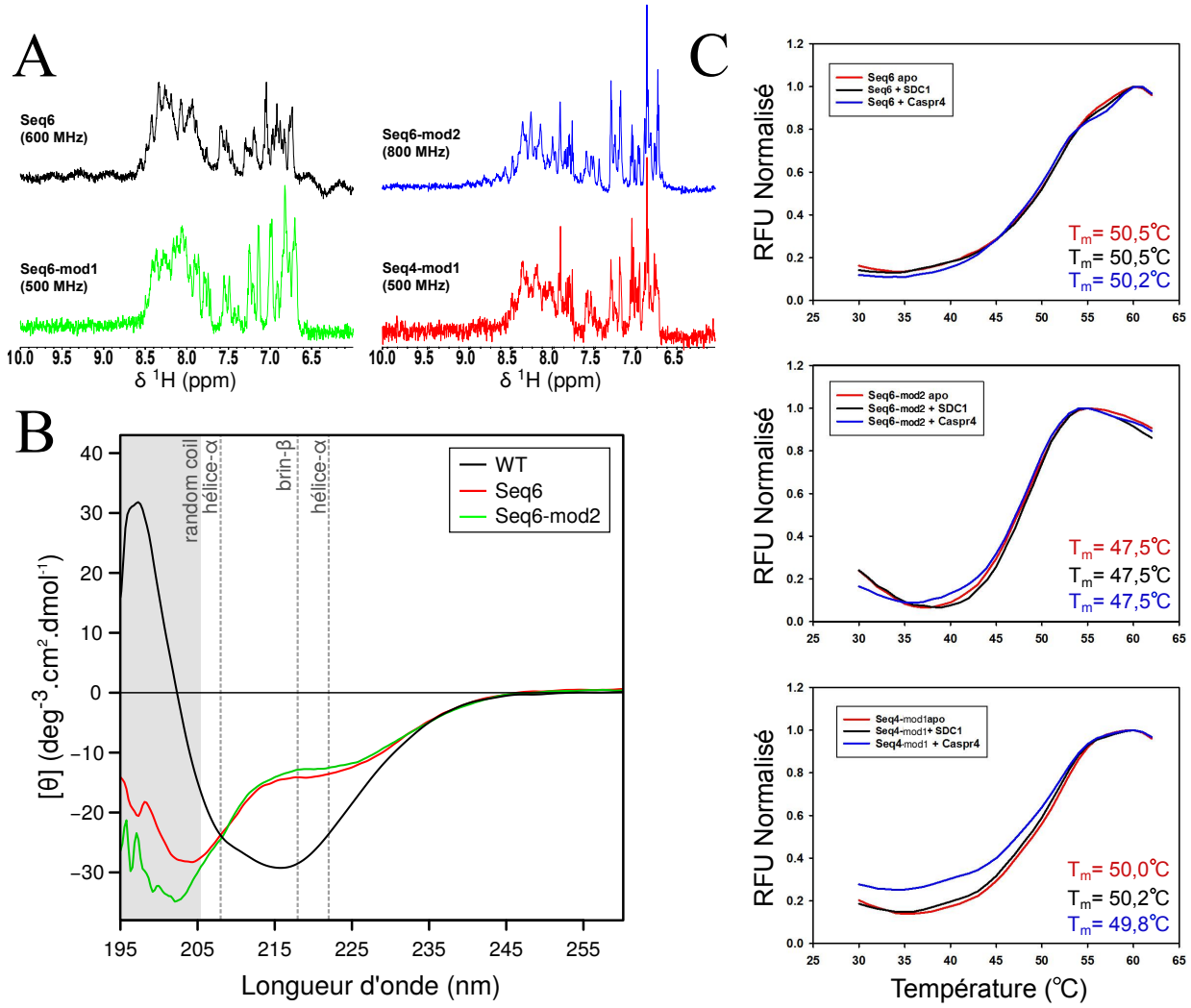


Figure 4.10 – Résultats expérimentaux des séquences Proteus. A : Spectre RMN 1-D. B : Estimation des structures secondaires par dichroïsme circulaire. C : Résultats de fluorimétrie différentielle à balayage en présence ou non des peptides Sdc1 et Caspr4. Toutes les analyses ont été effectuées dans un tampon phosphate (20 mM phosphate, 50 mM NaCl, 1 mM EDTA, pH 6.8). La concentration en protéines pour les expériences de RMN et dichroïsme circulaire en comprise entre 20 et 25 μM.

4.5.2 Modification manuelle des séquences

Les premiers résultats expérimentaux de la séquences 6 étant encourageants, nous avons identifié quelques positions dont la modification pourrait augmenter la stabilité. Pour cela, les séquences 4 et 6 ont été comparées à la séquence native et à une partie de l'alignement Pfam afin d'identifier des positions conservées et donc potentiellement importantes pour la stabilité du domaine. Nous nous sommes également appuyés sur les observations faites lors des simulations. Les positions identifiées sont mutées vers le type sauvage. Huit positions

pouvant stabiliser la protéine ont été identifiées. En localisant la position de ces régions sur la structure cristallographique quatre régions ont finalement été retenues :

Boucle β_1-β_2 Comme l'ont montré les simulations, les charges présentes dans la boucle β_1-β_2 augmentent sa flexibilité et peuvent partiellement masquer le site de fixation du peptide. Pour limiter la charge, les positions 853-855 (KEE/KER) sont remises à leur type natif (TAA).

Position 864 La présence d'une asparagine ou d'un glutamate à la position 864 déforme l'hélice α_2 en interagissant avec les sérines S907 et S908. Le retour au type natif permettrait donc de stabiliser l'hélice, ce qui pourrait également favoriser la liaison du peptide.

Position 893 Le glutamate à cette position est très conservé dans l'alignement Pfam. Cette conservation pourrait s'expliquer soit par un rôle fonctionnel, soit un rôle structural. Dans le cas où son rôle serait structural, il serait important de conserver ce résidu. Pour cela, les structures cristallographiques de trois domaines PDZ (4GVD, 2BYG et 1G9O) ont été analysées. Dans ces trois structures, la chaine latérale du glutamate interagit avec le squelette de la position 890 ce qui pourrait stabiliser le coude α_1-β_4. Cela suggère donc un rôle structural potentiel de E893.

Position 899 Tout comme E893, la position 899 présente une asparagine très conservée dans l'alignement Pfam. L'étude des trois mêmes domaines PDZ montre la présence d'interactions entre ce résidu et le squelette de la position 921 ce qui pourrait stabiliser le coude β_4-α_2.

Modifications supplémentaires En se basant sur leurs connaissances des domaines PDZ, l'équipe de E. Fuentes a également proposé quelques modifications supplémentaires pour la séquence 6. Ces mutations se situent notamment au niveau de la position 875 (retour au type natif V), de l'hélice α_1 (positions 883, 885 et 886), de la région 897 à 903, de l'hélice α_2 (position 914), de la boucle α_2-β_6 en ajoutant deux serines et du brin β_6 (positions 922 et 924).

Afin d'identifier l'impact de ces modifications sur la stabilité des séquences il serait intéressant de tester différentes combinaisons de mutations. Malheureusement, pour des raisons de coût et de temps, seules les séquences 4 et 6 présentant toutes les mutations ont pu être

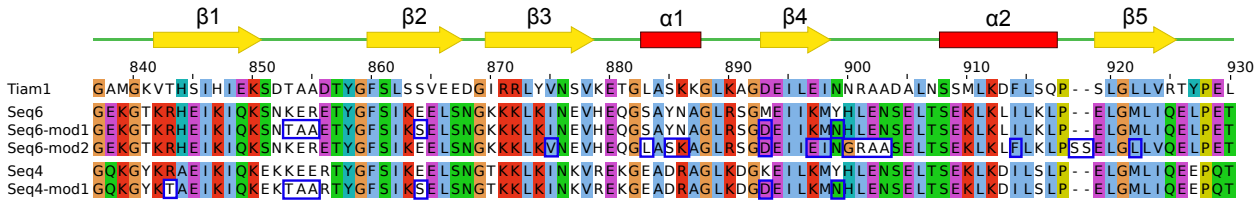


Figure 4.11 – Séquences modifiées manuellement. Les positions modifiées sont encadrées en bleu. Les structures secondaires sont indiquées au-dessus des séquences.

testées (seq4-mod1 et seq6-mod1). Les mutations proposées par l'équipe de E. Fuentes sur la séquence 6 (seq6-mod2) ont été appliquées séparément (figure 4.11). De nouveaux tests *in vitro* ont ainsi été effectués sur trois nouvelles séquences.

4.5.3 Analyses expérimentales des séquences modifiées

Les résultats obtenus pour les séquences sont assez proches de ceux de seq6 (figure 4.10A, B et C). Les séquences ont toutes tendance à former des multimères en solutions. Les résultats RMN indiquent que seq6-mod2 est légèrement plus structurée que les autres séquences. Les points de fusion restent inchangés et sont d'environ 50°C. Les modifications semblent donc augmenter la structuration de seq6-mod2. Le profil DC reste par ailleurs éloigné de celui de Tiam1. Les résultats obtenus restent encourageants car les séquences se replient partiellement et sont stables alors que 80% de la séquence a été modifiée.

4.6 Conclusions

Nous avons testé la stabilité des séquences Proteus par simulation de dynamique moléculaire. Pour cela, dix séquences ont été sélectionnées parmi les milliers de séquences générées par Proteus en se basant sur des critères physico-chimiques. Les dix séquences sélectionnées conservent toutes leur repliement lors les simulations avec cependant des fluctuations plus ou moins importantes selon les séquences. Malgré cette stabilité apparente, les séquences ne se replient pas complètement *in vitro* lors des premiers tests. Les simulations ont permis d'identifier quelques régions dont la modification pouvait améliorer la stabilité ce qui *in vitro* a été le cas. Ce repliement partiel est en soi un résultat prometteur puisque la stabilité des domaines PDZ semble extrêmement sensible aux modifications apportées à la séquence. Cela se voit particulièrement dans le cas du quadruple mutant où quatre mutations au niveau de l'hélice

α_2 suffisent à réduire de 1,6 kcal/mol la stabilité, rendant ce mutant déplié 20% du temps à température ambiante.

La comparaison de la dynamique aux données RMN montre une surestimation de la flexibilité des boucles et de l'hélice α_2 au cours des simulations. Malgré cette différence, les dynamiques permettent de reproduire les principales observations expérimentales, notamment en ce qui concerne la stabilisation du domaine par la liaison du peptide.

Une des principales limites de notre test provient de la longueur limitée des simulations. Ainsi, toutes les composantes principales ne sont pas parfaitement échantillonnées. Avec plus de ressources, on pourrait également simuler le dépliement thermique des différents variants et prédire la température de dénaturation.

Troisième partie

Estimation de l'affinité des complexes Tiam1-peptide par des modèles d'énergie libre semi-empiriques et exacts

Chapitre 5

Les méthodes de calcul d'énergie libre de liaison

Estimer l'affinité d'un complexe par des approches computationnelles représente l'un des défis majeurs de la modélisation moléculaire. En effet, l'association non covalente de biomolécules, comme une enzyme et son substrat, régit la majorité des processus cellulaires. L'interaction entre un récepteur et son ligand est un phénomène complexe pouvant mettre en jeu de nombreux effets : un changement dans l'entropie conformationnelle, translationnelle et vibrationnelle des partenaires, un réarrangement du solvant, la modification des interactions électrostatiques et de van der Waals entre les partenaires et avec le solvant ou encore la réorganisation de contre-ions (Simonson [2015]). Cette complexité rend l'estimation de l'affinité difficile et couteuse. Pour répondre à ce problème, différentes méthodes ont été développées.

Les approches les plus simples et les moins couteuses permettent d'estimer l'affinité d'un grand nombre de ligands pour une cible. Pour arriver à de telles performances, les termes énergétiques sont simplifiés et souvent couplés à des valeurs statistiques regroupées dans des fonctions de score. Cependant, la simplicité de ces fonctions fait qu'il est souvent difficile de discerner deux ligands ayant une différence d'affinité inférieure à 1,5 kcal/mol (Genheden & Ryde [2015]).

Pour déterminer précisément une affinité, les méthodes basées sur la mécanique statistique, appelées perturbation d'énergie libre (ou FEP pour *Free energy Perturbation*), peuvent être utilisées. Bien que prédicitives, ces méthodes sont extrêmement couteuses car elles nécessitent de simuler les états initiaux et finaux de la réaction mais également des intermédiaires non physiques reliant ces deux états. Ces méthodes sont donc généralement appelées méthodes alchimiques.

Entre ces deux groupes de méthodes un autre groupe donnant des performances intermédiaires existe. Comme le FEP, ces méthodes sont basées sur l'échantillonnage des conformations mais uniquement des états initiaux et finaux. Elles sont de ce fait moins couteuses que le FEP et plus justes que les fonctions de score. Elles nécessitent une étape de paramétrisation et sont donc semi-empiriques. Dans ce chapitre nous nous intéresserons aux méthodes FEP et semi-empiriques.

5.1 Modèles exacts pour le calcul de l'énergie libre de liaison

Les méthodes exactes, basées sur la mécanique statistique donnent généralement les meilleures performances. Elles sont toutefois couteuses puisqu'elles nécessitent de simuler des états intermédiaires reliant les états initiaux et finaux. Elles ont cependant l'avantage de ne pas avoir de paramètres ajustables. Les premières applications aux protéines remontent aux années 1980 (Wong & McCammon [1986] ; Warshel *et al.* [1986]). Ces méthodes peuvent être divisées en deux catégories, les méthodes alchimiques, pour lesquels les états intermédiaires correspondent à des états non physiques et les méthodes géométriques pour lesquels le processus de liaison est explicitement simulé par le biais de contraintes.

5.1.1 Calcul d'énergie libre par transformation alchimique

5.1.1.1 Théorie de la mécanique statistique pour le calcul de l'énergie libre

Dans un ensemble canonique où le nombre de particules, la température et la pressions sont constants (NpT), l'énergie libre de Gibbs d'un système A correspond à (Simonson [2001]) :

$$G_A = -kT \ln Q_A(N,p,T) + C \tag{5.1}$$

où k est la constante de Boltzmann, T la température, C une constante qui dépend des vitesses et Q_A la partie indépendante des vitesses de la fonction de partition :

$$Q_A(N,p,T) = \int \exp\left[-\frac{U_A + pV}{kT} \right] dr^N dV \tag{5.2}$$

où l'on intègre sur toutes les conformations possibles du système et sur le volume V.

En pratique, l'énergie libre est rarement calculée et l'on s'intéresse plutôt à la différence d'énergie libre entre deux systèmes proches. Ces systèmes, ou états, peuvent correspondre aux états dissociés et associés d'un complexe ou à la forme sauvage et mutée d'une protéine. On notera par la suite ces deux états A et B. La différence d'énergie libre entre ces états, $\Delta G_{A\to B}$, s'écrit alors :

$$\Delta G_{A\to B} = G_B - G_A = -kT \ln \frac{Q_B}{Q_A} = -kT \ln \frac{\int \exp(-U_B/kT) dr^N}{\int \exp(-U_A/kT) dr^N} \tag{5.3}$$

Soit ΔU, le changement d'énergie potentielle associé à la transformation de A vers B, tel que $U_B = U_A + \Delta U$. L'équation (5.3) devient alors :

$$\begin{aligned} \Delta G_{A\to B} &= -kT \ln \frac{\int \exp(-U_A/kT) dr^N \exp(-\Delta U/kT) dr^N}{\int \exp(-U_A/kT) dr^N} \\ &= -kT \ln \left\langle \exp\left[-\frac{\Delta U}{kT} \right] \right\rangle_A \end{aligned} \tag{5.4}$$

Les crochets, $\langle \rangle_A$, indiquent une moyenne sur un ensemble conformationnel issu de l'état A obtenu par simulation de dynamique moléculaire ou Monte Carlo. Le calcul ci-dessus (5.4) suppose que les systèmes A et B contiennent le même nombre de particules (les fonctions U_A et U_B opèrent sur le même espace).

Les relations (5.3) et (5.4) sont à la base de la plupart des calculs de différence d'énergie libre (Straatsma & McCammon [1992] ; Gilson & Zhou [2007]). Elles ne sont cependant applicables que si les états A et B sont suffisamment proches, ce qui n'est pas toujours le cas. Afin de pallier le problème, la transformation peut être remplacée par une série de transformations non physiques (ou alchimiques) reliant les états A et B (Valleau & Card [1972]). Pour cela un paramètre de couplage λ est introduit (Beveridge & DiCapua [1989]), qui pondère les interactions électrostatiques et van der Waals des deux états. Le paramètre λ varie de 0 à 1 de telle sorte que pour une valeur donnée, on a :

$$U(\lambda) = (1-\lambda)U_A + \lambda U_B \tag{5.5}$$

En sommant les différences d'énergie entre les états intermédiaires, on obtient la différence d'énergie libre entre A et B. Différentes méthodes permettent de calculer la différence d'énergie

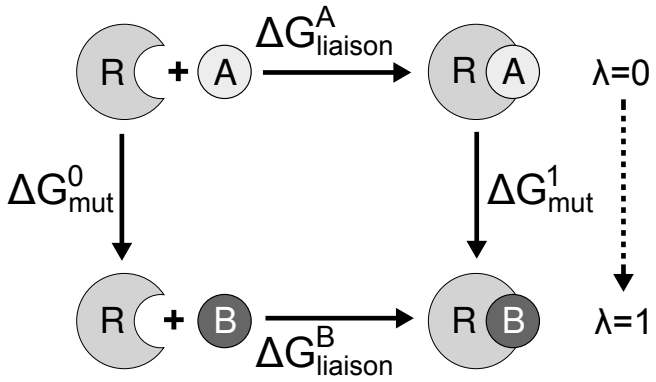


Figure 5.1 – Cycle thermodynamique permettant de calculer l'energie libre de liaison relative. A et B correspondent à deux ligands se fixant sur le même récepteur R. ΔG^0_{mut} et ΔG^1_{mut} correspondent respectivement aux énergies libres de mutations du ligand A vers B dans les états dissociés et associés. $\Delta G^A_{liaison}$ et $\Delta G^B_{liaison}$ correspondent aux énergies libres de liaison des ligands A et B respectivement.

entre les états intermédiaires, on peut notamment citer l'intégration thermodynamique (TI), le ratio d'acceptation de Bennett (BAR) ou encore la *Weighted Histogram Analysis Method* (WHAM) que nous détaillerons plus loin (partie 5.1.1.3).

5.1.1.2 Énergies libres absolues et relatives

Le FEP permet de déterminer la différence d'énergie libre entre plusieurs états qui peuvent correspondre à l'état associé et dissocié de plusieurs complexes protéine:ligand. Dans ce cas, on peut utiliser le cycle thermodynamique présenté en figure 5.1. Parfois, le but n'est pas de déterminer l'énergie libre de liaison d'un ligand mais de comparer plusieurs ligands. Il serait pour cela possible de calculer l'énergie libre absolue de chaque ligand, néanmoins une méthode plus efficace consiste à calculer l'énergie libre associée à la mutation du ligand A vers le ligand B dans les états associés et dissociés. Le cycle thermodynamique décrivant la réaction est présenté en figure 5.1. Les flèches horizontales ($\Delta G^A_{liaison}$ et $\Delta G^B_{liaison}$) correspondent à l'énergie de liaison des ligands A et B au récepteur R. L'énergie libre de liaison relative peut donc s'écrire :

$$\Delta\Delta G_b^{A\to B} = \Delta G^B_{liaison} - \Delta G^A_{liaison} \tag{5.6}$$

Les flèches verticales correspondent aux énergies libres associées à la transformation alchimique du ligand A vers B dans son état dissocié (ΔG^0_{mut}) et lié au récepteur (ΔG^1_{mut}). Ces quatre

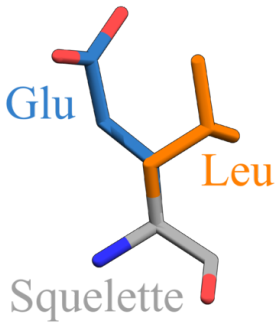


Figure 5.2 – Exemple d'une topologie double. Les chaines latérales des résidus Leu (orange) et Glu (bleu) sont positionnées sur le même squelette (gris) et n'interagissent pas entre elles.

étapes formant un cycle thermodynamique, la somme des valeurs de ΔG est nulle. On a donc :

$$\Delta G^0_{mut} + \Delta G^B_{liaison} - \Delta G^1_{mut} - \Delta G^A_{binding} = 0 \tag{5.7}$$

soit

$$\Delta\Delta G^{A\to B}_b = \Delta G^B_{liaison} - \Delta G^A_{liaison} = \Delta G^1_{mut} - \Delta G^0_{mut} \tag{5.8}$$

Ainsi, il est possible de calculer l'énergie libre de liaison relative associée à la mutation $A \to B$ à partir de la différence d'énergie libre liée à la transformation alchimique dans les formes associée et dissociée du complexe (Straatsma & McCammon [1991]).

Lorsque les ligands ne diffèrent que par un groupement chimique présent dans l'un et absente dans l'autre, il est généralement suffisant de découpler progressivement ce groupement. Au contraire, lorsque les ligands sont très différents (et c'est notamment le cas lorsqu'il s'agit de muter un acide aminé vers un autre type), il est nécessaire d'utiliser le paradigme de la topologie double (Tembe & McCammon [1984] ; Jorgensen & Ravimohan [1985]). Dans le cas d'un acide aminé, la topologie double peut correspondre à un squelette unique sur lequel deux chaines latérales sont greffées au niveau du C_α (figure 5.2). Ces deux chaines latérales coexistent au cours de la transformation mais n'interagissent pas entre elles. Quand $\lambda = 0$ (respectivement $\lambda = 1$), le mutant (respectivement le sauvage) est découplé du reste du système et ne conserve que les interactions avec lui-même.

Pour calculer l'énergie libre d'association d'un complexe, on peut utiliser le cycle présenté en figure 5.3. Dans ce cycle, l'énergie libre de liaison correspond à la flèche horizontale supérieure et représente l'association du ligand et du récepteur. On peut donc calculer l'énergie

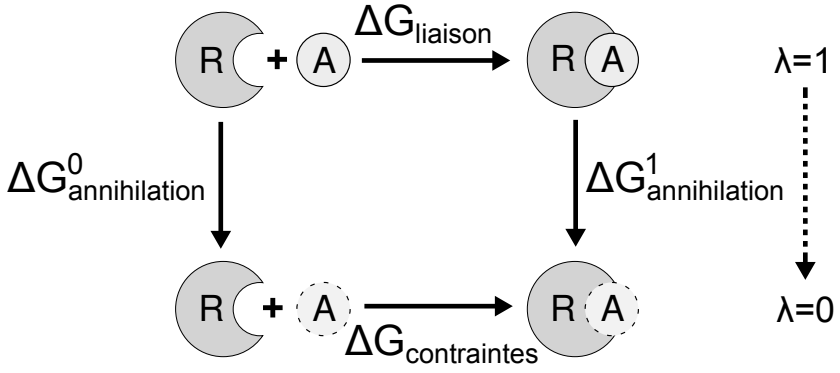


Figure 5.3 – Cycle thermodynamique permettant de calculer l'énergie libre de liaison absolue.. La flèche horizontale supérieure représente l'énergie libre de liaison du ligand A au récepteur B. $\Delta G^0_{annihilation}$ et $\Delta G^1_{annihilation}$ correspondent respectivement aux énergies libres associées au découplage du ligand dans les états dissocié et associé. $\Delta G^0_{contraintes}$ représente la perte d'entropie due à la contrainte nécessaire pour que le ligand adopte sa conformation liée.

libre de liaison à partir des énergies libres associées au découplage du ligand dans son état dissocié ($\Delta G^0_{annihilation}$) et associé ($\Delta G^1_{annihilation}$). Dans sa forme libre, le ligand explore un espace conformationnel beaucoup plus grand que lorsqu'il est lié. La perte d'entropie conformationnelle associée à la liaison du ligand ne doit donc pas être négligée lors du calcul d'énergie libre absolue. Elle est ici représentée par le terme $\Delta G_{contraintes}$ et correspond généralement à l'énergie de contrainte nécessaire pour maintenir le ligand découplé dans sa conformation liée.

5.1.1.3 Estimation de l'énergie libre

Trouver le meilleur estimateur de la différence d'énergie étant donnés deux ensembles de configurations n'est pas un problème trivial. Nous présentons ici quelques-unes des principales méthodes couramment utilisées.

Méthode exponentielle La méthode exponentielle est l'une des plus anciennes méthodes permettant de calculer la différence d'énergie libre entre deux états. Initialement proposée par Peierls [1933] et Landau & Lifshitz [1951], puis Zwanzig [1954], cette approche correspond à l'équation (5.4) et permet de calculer la différence d'énergie libre entre deux états. La valeur du $\Delta G_{0\to1}$ est calculée en sommant les différences d'énergies libres intermédiaires :

$$\Delta G_{0\to1} = \sum_{i=0}^{N-1} -kT \ln \left\langle \exp\left[-\frac{U(\lambda_{i+1}) - U(\lambda_i)}{kT} \right] \right\rangle_{\lambda_i} \tag{5.9}$$

avec N le nombre de valeurs de λ intermédiaires et $\langle\rangle$ représentant une moyenne sur un ensemble conformationnel. Cette méthode est très simple mais converge lentement (Shirts & Pande [2005]).

Intégration thermodynamique Dans l'intégration thermodynamique (TI), l'énergie libre de mutation est décrite comme une somme d'intégrales sur λ :

$$\Delta G_{0\to 1} = \int_0^1 \frac{\partial G}{\partial \lambda} d\lambda = \sum_{i=0}^{n-1} \int_{\lambda_i}^{\lambda_{i+1}} \frac{\partial G}{\partial \lambda} d\lambda \tag{5.10}$$

La somme est effectuée sur n intervalles $[\lambda_i,\lambda_{i+1}]$, avec $\lambda_0 = 0$ et $\lambda_n = 1$. Pour calculer numériquement l'intégrale, la méthode la plus simple est celle des trapèzes :

$$\sum_{i=0}^{n-1} \int_{\lambda_i}^{\lambda_{i+1}} \frac{\partial G}{\partial \lambda} d\lambda \approx \frac{1}{2}(\lambda_{i+1} - \lambda_i)\left(\frac{\partial G}{\partial \lambda}(\lambda_i) + \frac{\partial G}{\partial \lambda}(\lambda_{i+1})\right) \tag{5.11}$$

Cette méthode requiert généralement plus de valeurs intermédiaires de λ et n'a pas d'aussi bonnes performances que la méthode BAR (ci-dessous) (Shirts & Pande [2005]). D'autres méthodes ont donc été développées. Parmi celles-ci, on peut citer l'approximation des *splines* cubiques qui interpole les valeurs de $\frac{\partial G}{\partial \lambda}(\lambda)$ par une fonction polynomiale et qui est largement utilisée. D'autres méthodes comme la loi de Simpson, la quadrature de Gauss-Legendre ou l'intégration de Clenshaw-Curtis (Bruckner & Boresch [2010]) sont plus rarement appliquées.

Ratio d'acceptation de Bennett Le ratio d'acceptation de Bennett est un estimateur de l'énergie libre qui minimise la variance (Bennett [1976]). Bennett a ainsi montré que la différence d'énergie libre entre les deux états i et j (correspondant dans notre cas à deux valeurs de λ adjacentes) peut être calculée de la manière suivante :

$$\Delta G_{ij}^{BAR} = k_B T(ln\frac{\langle f(\Delta U + C)\rangle_j}{\langle f(-\Delta U - C)\rangle_i}) + C \tag{5.12}$$

avec f la fonction de Fermi

$$f(x) = \frac{1}{1 + exp(\frac{x}{k_B T})} \tag{5.13}$$

avec ΔU la différence d'énergie entre les états i et j. La valeur de C est déterminée de manière itérative jusqu'à vérifier :

$$f(\Delta U + C) = f(-\Delta U - C) \tag{5.14}$$

La différence d'énergie est alors calculée comme suit (de Ruiter *et al.* [2013]) :

$$\Delta G_{ij}^{BAR} = -k_B T \ln \frac{N_j}{N_i} + C \tag{5.15}$$

avec N_i et N_j le nombre de conformations à λ_i et λ_j respectivement. Si $N_i = N_j$, on obtient :

$$\Delta G_{ij}^{BAR} = C \tag{5.16}$$

La méthode BAR donne des résultats comparables aux autres méthodes telles que l'intégration thermodynamique tout en étant moins sensible au choix des intermédiaires, ce qui la rend plus robuste (Paliwal & Shirts [2011] ; de Ruiter *et al.* [2013]). Une variante du BAR, appelée ratio d'acceptation de Bennett multi-états (MBAR) estime la différence d'énergie libre à partir de l'échantillonnage de plus de deux états (Shirts & Chodera [2008]).

Weighted Histogram Analysis Method La *Weighted Histogram Analysis Method* (WHAM) fut développée par Ferrenberg & Swendsen [1989] puis adaptée aux transformations alchimiques par Kumar *et al.* [1992]. Comme MBAR, le WHAM estime la différence d'énergie libre en prenant en compte tous les états intermédiaires. Le WHAM nécessite de discrétiser l'espace des états pour pouvoir créer un histogramme dont les barres représentent la probabilité d'observer un état donné le long de la transformation. À partir de ces probabilités, on peut estimer la différence d'énergie libre entre les deux états.

5.1.2 Calcul d'énergie libre par transformation géométrique

La constante d'équilibre, K_{eq}, peut être reliée à un potentiel de force moyenne, $w(r)$, où r est une séparation du ligand et du récepteur (Shoup & Szabo [1982]) :

$$K_{eq} = 4\pi \int_0^R r^2 e^{-\beta w(r)} dr \tag{5.17}$$

R représente la limite de l'association et $w(r)$ est l'énergie libre de séparation. Cette approche suppose cependant que le ligand échantillonne l'ensemble de l'espace conformationnel au cours de la simulation, ce qui est rarement le cas. Afin de pallier le problème, une possibilité consiste à décomposer le processus de liaison en une série de transformations géométriques qui contraignent progressivement le ligand vers la forme liée, puis dans un second temps à

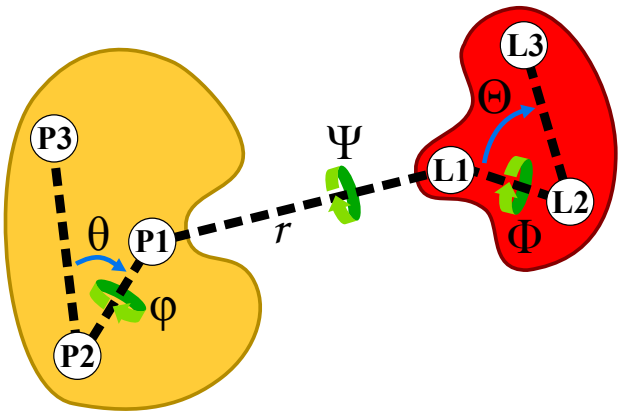


Figure 5.4 – Représentation schématique des données de références permettant de définir la position du ligand par rapport au récepteur et de construire les potentiels de contraintes. Pour la protéine et le ligand, trois groupes d'atomes sont choisis et froment les triplets P1,P2,P3 et L1,L2,L3 respectivement. La position du ligand par rapport au récepteur est déterminée par la valeur moyenne des coordonnées sphériques r,θ,ϕ où r correspond à la distance P1-L1, θ à l'angle L1-P1-P2 et ϕ à l'angle L1-P1-P2-P3. L'orientation relative du ligand est décrit par les trois angles d'Euler $\Theta(P1-L1-L2)$,$\phi(P1-L1-L2-:3)$,$\Psi(P2-P1-L1-L2)$. Figure adaptée de Woo & Roux [2005]

calculer l'énergie libre associée à ces contraintes (Woo & Roux [2005] ; Gumbart *et al.* [2013]). Le processus de liaison étant explicitement simulé, la méthode géométrique calcule l'énergie libre de liaison absolue.

Pour pousser le ligand vers son état lié, il faut définir sa position par rapport au récepteur. On peut choisir pour cela trois atomes dans la protéine et dans le ligand et introduire les coordonnées sphériques (r,θ,ϕ) et les angles d'Euler (Θ,Φ,Ψ) du ligand par rapport au récepteur (voir figure 5.4). À ces descripteurs s'ajoutent deux potentiels supplémentaires : un potentiel qui contraint le ligand sur un axe orienté vers le site de liaison et un potentiel qui maintient le ligand dans une conformation proche de celle de l'état lié. En agissant sur ces potentiels, on peut contraindre progressivement le ligand vers sa forme liée. L'énergie libre associée à chaque contrainte peut être estimée par l'approche de l'*Adaptive biasing force* (ABF, Comer *et al.* [2015]) ou par *umbrella sampling*. Afin d'améliorer la convergence des résultats, les coordonnées de réaction sont la plupart du temps divisées en fenêtres intermédiaires.

La méthode géométrique est généralement privilégiée lorsque l'énergie de solvatation du ligand est élevée car, dans ce cas, les méthodes FEP introduisent une incertitude importante (Woo & Roux [2005]). Cette approche a notamment été appliquée avec succès dans le calcul

de l'énergie libre de liaison d'un peptide phosphorylé au domaine SH3 de la protéine Abl avec une erreur de 0,3 kcal/mol par rapport à l'énergie libre expérimentale (Woo & Roux [2005]).

5.2 Modèles semi-empiriques pour l'énergie libre

Les méthodes exactes sont prédictives mais couteuses et difficilement applicables à un grand nombre de ligands. Des approches moins rigoureuses et moins couteuses, qui ne simulent que les états initiaux et finaux de la réaction, sont possibles. Le processus de liaison peut être décomposé en plusieurs étapes décrites par le cycle thermodynamique présenté en figure 5.5. Les charges de la forme dissociée sont d'abord supprimées, laissant un soluté apolaire ; puis les interactions de dispersion sont supprimées et les deux partenaires s'associent ; puis les interactions de dispersion et les charges sont restaurées. L'énergie libre de liaison, ΔG_b, est ensuite déterminée de la manière suivante :

$$\Delta G_b = \langle G_{associé} \rangle - \langle G_{dissocié} \rangle \tag{5.18}$$

où $G_{associé}$ représente l'énergie libre de la forme associée du complexe et $G_{dissocié}$ celle de la forme dissociée. La plupart du temps, les énergies libres sont moyennées sur un ensemble de conformations plutôt que calculées sur une conformation unique. Pour cela, des simulations de dynamique moléculaire ou Monte Carlo sont effectuées puis des conformations sont extraites à intervalles réguliers. Cet ensemble conformationnel est symbolisé dans l'équation (5.18) par les crochets.

Parmi les méthodes basées sur cette approximation, nous pouvons citer les modèles s'appuyant sur l'approximation de la réponse linéaire, les modèles d'énergies d'interaction linéaire ou encore les modèles de types MM/PBSA. De nombreux variants de ces méthodes ont été développés au cours des dernières années, nous ne présenterons ici que leurs caractéristiques principales.

5.2.1 Les modèles d'énergies d'interaction linéaire

Plusieurs méthodes sont basées sur l'approximation de la réponse linéaire (ou LRA pour *Linear Response Approximation*). Ainsi, Åqvist *et al.* modélisèrent les interactions électrostatiques et apolaires par les termes de Coulomb et van der Waals (Åqvist *et al.* [1994, 2002])

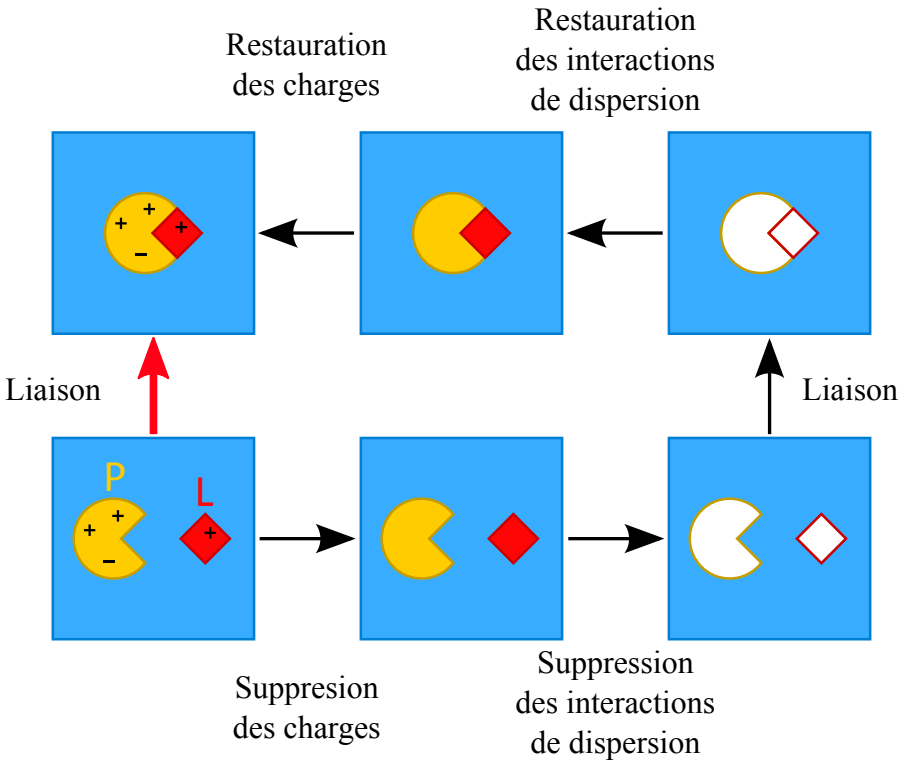


Figure 5.5 – Cycle thermodynamique du processus de liaison. La flèche verticale à gauche représente la liaison de ligand L à la protéine P. Le chemin vers la droite correspond aux étapes suivantes : (1) les charges de la forme dissociée sont tout d'abord supprimées, laissant un soluté apolaire ; (2) les interactions de dispersion sont ensuite à leur tour supprimées ; (3) les deux partenaires s'associent ensuite ; (4) puis les interactions de dispersion sont restaurées ; (5) les charges sont à leur tour restaurées.

dans un modèle appelé LIE (pour *Linear Interaction Energy*). Le LIE nécessite de simuler le ligand dans ses formes libre et associée au récepteur. L'énergie libre de liaison est calculée de la manière suivante :

$$\Delta G_b(l) = \alpha(\langle U_{PL}^{vdW} \rangle - \langle U_L^{vdW} \rangle) + \beta(\langle U_{PL}^{elec} \rangle - \langle U_L^{elec} \rangle) + \gamma \tag{5.19}$$

Les crochets représentent une moyenne sur un ensemble de conformations. Les coefficients α, β et γ sont des constantes empiriques, ajustées afin de reproduire les affinités expérimentales (Wang *et al.* [1999]). Le terme γ s'annule lors du calcul d'énergies libres de liaison relatives. Le coefficient β est généralement fixé à une valeur de 0,5 d'après la théorie de la réponse linéaire appliquée aux forces électrostatiques (Marcus [1964]). D'autres études proposent des valeurs différentes de β en fonction de la nature chimique du composé (Åqvist & Hansson [1996] ; Almlöf *et al.* [2007]).

5.2.1.1 Exemples d'applications des modèles LIE

Les modèles LIE ont été largement appliqués, notamment dans les cas de la dihydrofolate réductase (Graffner-Nordberg *et al.* [2001]), de la thrombine (Kajsa B. Ljungberg *et al.* [2001]), la neuraminidase (Wall *et al.* [1999]) et plus récemment la phosphodiestérase 10A (Kjellgren *et al.* [2015]). Dans certains cas, un terme surfacique est également ajouté pour compléter la description des interactions apolaires (Smith *et al.* [1998] ; Lamb *et al.* [1999]). Dans toutes ces études les ligands sont de petits composés chimiques. Cette méthode permet généralement d'obtenir des valeurs d'erreurs quadratiques moyennes entre les valeurs expérimentales et calculées d'environ 1 kcal/mol (Åqvist & Marelius [2001] ; Gutierrez-de Teran & Åqvist [2011]).

5.2.2 Les modèles MM/PBSA et modèles apparentés

La méthode MM/PBSA est certainement la plus utilisée. Initialement développée par Kollman *et al.* (Srinivasan *et al.* [1998] ; Kollman *et al.* [2000]), elle estime l'énergie libre de liaison à partir de l'énergie libre des partenaires associés et dissociés :

$$\Delta G_b = \langle G_{PL} \rangle - \langle G_P \rangle - \langle G_L \rangle \tag{5.20}$$

L'énergie libre de chaque état est calculée comme suit :

$$G = E_{MM} + G_{solv} - TS \tag{5.21}$$

où E_{MM} correspond à l'énergie totale calculée à l'aide de la mécanique moléculaire, G_{solv} représente l'énergie libre de solvatation, S est l'entropie conformationnelle et T la température. Comme dans le cas du LIE, les approches de type MM/PBSA sont généralement appliquées à un ensemble de conformations générées par dynamique moléculaire ou Monte Carlo en solvant explicite. Les différents termes énergétiques sont ensuite déterminés lors d'une étape de post-traitement pendant laquelle les molécules d'eau sont retirées et remplacées par un continuum diélectrique. Nous allons maintenant détailler plus précisément les différents termes énergétiques.

5.2.2.1 Description des termes énergétiques

Termes de mécanique moléculaire Le terme E_{MM} est issu de la mécanique moléculaire et décrit les interactions liées (liaisons, angles, dièdres, impropres) et non liées entre les atomes (électrostatique et van der Waals) du soluté. Elles sont calculées à partir des paramètres du champ de force.

$$E_{MM} = E_{liée} + E_{elec} + E_{vdW} \tag{5.22}$$

Les termes d'énergie liée de la protéine et du complexe peuvent être de l'ordre de quelques centaines de kcal/mol et introduire une incertitude importante dans l'estimation de l'énergie libre. Afin de limiter cette incertitude les conformations des états lié et dissocié sont généralement extraites de la trajectoire du complexe (Gohlke & Case [2004] ; Lepsik *et al.* [2004] ; Genheden & Ryde [2012]). De ce fait, le récepteur et le ligand présentent exactement la même conformation dans les deux états, les termes $E_{liée}$ s'annulent donc. Cette approche, appelée mono-trajectoire, réduit considérablement le temps de calcul puisqu'il n'est plus nécessaire de simuler le récepteur et le ligand dans leur état dissocié. L'approche mono-trajectoire fait cependant l'hypothèse que la liaison n'entraine pas de changements conformationnels importants. Il est toutefois possible de décrire implicitement la réorganisation des charges en utilisant une constante diélectrique supérieure à un pour les solutés (Schutz & Warshel [2001]). Une autre approche consiste à simuler le ligand dans son état dissocié pour prendre en compte son énergie de réorganisation (Swanson *et al.* [2004] ; Yang *et al.* [2009]).

Énergie de solvatation L'énergie de solvatation, E_{solv}, décrit les interactions entre le soluté et le solvant mais également l'écrantage des interactions électrostatiques entre les atomes du soluté. Actuellement, le meilleur moyen de décrire le solvant consiste à le traiter de manière explicite à l'aide de modèles moléculaires représentant les molécules d'eau (par exemples les modèles SPC, SPC/E, TIP3P, TIP4P ou TIP5P). Ces modèles sont cependant trop couteux pour permettre un calcul rapide. Les modèles MM/PBSA traitent donc le solvant de manière implicite. L'énergie de solvatation est alors décomposée en deux termes, un terme G^{pol}_{solv} décrivant les effets électrostatiques du solvant et un terme G^{apol}_{solv} décrivant les interactions apolaires entre le soluté et le solvant.

$$G_{solv} = G^{pol}_{solv} + G^{apol}_{solv} \tag{5.23}$$

Le terme G_{solv}^{pol} est généralement traité par un terme de Poisson-Boltzmann (PB) ou de Born généralisé (GB) dans lesquels le soluté forme une cavité de constante diélectrique ϵ_P faible (généralement comprise entre 1 et 8) entourée d'un continuum de constante diélectrique ϵ_W élevée (généralement 80). Le terme G_{solv}^{apol} correspond à la formation d'une cavité dans le solvant ainsi qu'aux forces attractives et répulsives de van der Waals entre le soluté et le solvant. Il est généralement proportionnel à la surface exposée au solvant (SASA).

Terme entropique Le terme TS vise à prendre en compte l'entropie conformationnelle du soluté. Ce terme est généralement calculé par une analyse en modes normaux de conformations issues de la simulation puis minimisées. Cette approche est cependant couteuse car un grand nombre de conformations est généralement nécessaire pour que les valeurs convergent (Genheden & Ryde [2012]). On peut également adopter une approximation quasi-harmonique. Le caractère prédictif de ce terme n'a pas été démontré ou très rarement. Il est généralement omis lors des calculs d'affinité, notamment dans les calculs d'énergies libres relatives, lorsque les deux ligands sont très similaires et se lient de la même façon à la protéine (Foloppe & Hubbard [2006] ; Rastelli *et al.* [2010]).

5.2.2.2 Exemples d'application des modèles MM/PB(GB)SA

Les modèles MM/PB(GB)SA ont souvent été utilisés ces dernières années pour pour caractériser des petits ligands (Kuhn & Kollman [2000] ; Hou *et al.* [2002] ; Stoica *et al.* [2008]) parfois liés à de l'ARN (Gouda *et al.* [2002]), des peptides (Venken *et al.* [2011] ; Spiliotopoulos *et al.* [2012]), notamment liés aux domaines de type SH3 (Hou *et al.* [2006]) ou encore des interactions protéine-protéine (Wang & Kollman [2000]). Comme on peut le voir dans le tableau 5.1, les résultats sont généralement corrélés avec les valeurs expérimentales mais les erreurs moyennes sont souvent très importantes (Hou *et al.* [2011] ; Wan *et al.* [2015]). Ces modèles ne donnent donc qu'une information qualitative sur les énergies libres de liaison.

5.2.3 Modèles PB/LIE et modèles apparentés

Dans les modèles de type MM/PB(GB)SA, les termes d'énergie libre issus de la mécanique moléculaire et les termes de solvatation ne sont généralement pas pondérés par des coefficients. Les résultats obtenus sont donc la plupart du temps plus qualitatifs que quantitatifs. Il est cependant possible d'estimer l'énergie libre de liaison en pondérant, à la manière de la méthode

LIE, les différents termes énergétiques issus du MM/PBSA. Nous ferons par la suite référence à cette méthode par le nom PB/LIE. L'énergie libre de liaison est calculée de la manière suivante :

$$\Delta G_b = \alpha \Delta G_{\text{elec}} + \beta \Delta E_{\text{vdW}} + \gamma \Delta S + \delta \tag{5.24}$$

où α, β, γ et δ sont des paramètres ajustables. La constante δ disparait lors du calcul de l'énergie libre de liaison relative. Comme dans le cas du MM/PBSA, le terme PB peut être remplacé par un terme de GB (GB/LIE). Comme pour le LIE, cette méthode nécessite une étape préalable consistant à optimiser les coefficients à partir d'un jeu d'affinités expérimentales.

Tableau 5.1 – Performances de quelques modèles MM/PBSA ou MM/GBSA.

Référence	Cible	Taille échant.	Modèle élec	ϵ_P	Termes pondérés	Estimation entropie[a]	Intervalle $\Delta\Delta G$	Performances modèle MUE	RMSD	R	Modèle Nul[b] MUE	RMSD
petites molécules												
Kuhn & Kollman [2000]	avidine	9	PB	1	SA	NM	[-5,0 ;-20,4]	3.3	3,8	0,96	4,1	4,9
Wang & Kollman [2000]	HIV-I protéase	12	PB	?	SA	NM	[-11,9 ;-7,8]	1,1	1,2	0,74	1,1	1,3
Rastelli *et al.* [2010]	DHFR	22	PB	1	SA	NM	[-15,0 ;-5,5]	7,2	8,0	0,91	2,4	2,7
Rastelli *et al.* [2010]	DHFR	22	GB	1	SA	NM	[-15,0 ;-5,5]	16,4	18,0	0,94	2,4	2,7
Hou *et al.* [2011]	$\alpha - thrombine$	7	PB	4	SA	NM	[-12,4 ;-4,0]	19,8	20,7	0,80	2,1	2,6
Hou *et al.* [2011]	avidine	7	PB	1	SA	NM	[-20,4 ;-4,5]	7,8	8,1	0,93	4,7	5,3
Hou *et al.* [2011]	cytochrome-C	18	PB	1	SA	NM	[-7,1 ;-3,8]	11,9	12,4	0,30	0,7	0,8
Hou *et al.* [2011]	Neuraminidase	8	PB	4	SA	NM	[-11,5 ;-3,7]	2,3	2,9	0,68	2,2	2,6
Hou *et al.* [2011]	P450cam	9	PB	1	SA	NM	[-7,9 ;-5,5]	5,8	6,1	0,68	0,8	0,9
Hou *et al.* [2011]	penicillopepsine	7	PB	2	SA	NM	[-12,8 ;-6,8]	2,8	3,0	0,41	2,1	2,3
Ben-Shalom *et al.* [2017]	HIV-I protéase	20	PB	1	Elec, vdW, SA, TS	NM	[-16,5 ;-9,9]	1,0	1,2	0,72	1,8	2,0
Ben-Shalom *et al.* [2017]	FXa	20	PB	1	Elec, vdW, SA, TS	NM	[-12,8 ;-8,6]	0,6	0,8	0,54	0,8	1,5
Ben-Shalom *et al.* [2017]	Hsp90	16	PB	1	Elec, vdW, SA, TS	NM	[-10,5 ;-5,1]	1,2	1,4	0,63	1,6	1,8
molécules peptidomimétiques												
Hou *et al.* [2002]	MMP-2	8	PB	1	Elec, vdW, SA	NM	[-15,0 ;-7,8]	3,0	3,6	0,84	1,9	2,2
ARN:bases azotées												
Gouda *et al.* [2002]	aptamère d'ARN	5	PB	1	SA	-	[1,1 ;5,6]	3,5	3,8	0,66	1,2	1,5
Hu *et al.* [2017]	aptamère d'ARN	6	PB	?	SA	NM	[-11,5 ;-6,0]	6,5	6,6	0,75	1,2	1,7
protéine:peptide												
Hou *et al.* [2006]	Abl-SH3	14	PB	1	SA	NM	[-0,8 ;2,6]	2,0	2,3	0,82	0,8	0,9
Stoica *et al.* [2008]	HIV-I protéase	4	PB	1	SA	NM	[-13,8 ;-10,6]	0,5	0,6	0,96	1,0	1,2
Venken *et al.* [2011]	HIV-1 gp41	29	PB	var.	SA	QH	[-2,4 ;1,1]	-	-	0.71	-	-
Spiliotopoulos *et al.* [2012]	AIRE-PHD1	9	PB	1	SA	NM	[0,2 ;2,1]	6,7	7,7	0,74	0,5	0,6

[a] : Méthode d'estimation de l'entropie conformationnelle utilisée : Modes normaux (NM), approximation quasi-harmonique (QH).
[b] : Le modèle Nul correspond à un modèle pour lequel toutes les valeurs prédites ont la même affinité correspondant ici à la moyenne des valeurs expérimentales.

Chapitre 6

Mise en place de modèles semi-empiriques pour la prédiction de ligands peptidiques

Dans ce chapitre nous nous intéressons à l'application de modèles semi-empiriques aux complexes Tiam1:peptide. Ces approches décrivent les interactions entre partenaires par une fonction d'énergie libre contenant un terme électrostatique utilisant un continuum diélectrique, un terme de surface accessible au solvant et un terme de van der Waals. Les modèles utilisés ici sont sensiblement différents des modèles MM/PBSA ou MM/GBSA classiques par l'introduction de coefficients pondérant les différents termes, ce qui les rapproche des modèles LIE. Nous les appellerons donc PB/LIE ou GB/LIE. L'optimisation de ces coefficients représente l'une des principales difficultés car elle nécessite un jeu de données expérimentales suffisamment conséquent et représentatif des ligands. Le domaine PDZ de la protéine Tiam1 est un parfait candidat pour tester ce type de méthodes puisque de nombreuses données expérimentales, aussi bien structurales que d'affinité, sont disponibles. À partir de ces données, 51 complexes ont étés modélisés puis simulés par dynamique moléculaire afin d'extraire les termes énergétiques nécessaires. La capacité des modèles à reproduire les affinités expérimentales constitue un bon test pour la classe des modèles PB(GB)SA. Ces modèles peuvent ensuite être utilisés à des fins prédictives pour identifier de nouveaux peptides liants ou de nouveaux variants de Tiam1. Enfin, la mise en regard des simulations et des affinités permet d'apporter des informations sur les mécanismes de reconnaissance des complexes PDZ:peptide.

6.1 Fonction d'énergie libre semi-empirique

Afin d'estimer l'énergie libre de liaison, ΔG_b, nous avons utilisé la fonction d'énergie libre suivante :

$$\Delta G_B = \alpha \Delta G_{elec} + \beta \Delta E_{vdW} + \gamma \Delta A + \delta \tag{6.1}$$

où α, β et γ sont des constantes ajustables. ΔG_{elec} représente la différence d'énergie libre électrostatique entre les états liés et non liés, calculée à l'aide d'un modèle PB ou GB et moyennée sur des conformations issues de simulations de dynamique moléculaire du complexe en solvant explicite. ΔE_{vdW} correspond aux interactions de van der Waals entre les atomes du peptide et de la protéine, moyennées sur des conformations issues de la trajectoire. ΔA est le changement de la surface moléculaire du soluté lors de la liaison et est donc négatif. Ce terme est également moyenné sur les conformations de la trajectoires. Nous avons fait le choix d'utiliser l'approche mono-trajectoire ce qui signifie que les conformations du peptide et de la protéine dissociées sont issues de la simulation du complexe. Le dernier terme, δ, est une constante qui disparait dans notre cas puisque nous calculons l'énergie libre de liaison relative par rapport au complexe Tiam1:Sdc1.

La fonction d'énergie utilisée est très proche des modèles MM/PB(GB)SA classiques mais contrairement à ces derniers, les différentes composantes énergétiques sont ici pondérées par des constantes. La présence de ces paramètres ajustables rend ce modèle également très proche des modèles de type LIE. Nous ferons donc par la suite référence à ces modèles par les termes PB/LIE et GB/LIE. Comme dans les modèles LIE, les modèles PB(GB)/LIE nécessitent une première étape de paramétrisation afin de déterminer les valeurs optimales des coefficients α, β et γ. Pour cela, il faut un jeu d'affinités expérimentales.

6.2 Données expérimentales disponibles

La mise en place de modèles semi-empiriques nécessite une première phase d'ajustement des paramètres sur un jeu d'apprentissage. Une fois le modèle optimisé, il peut être utilisé pour prédire l'affinité de nouveaux peptides. De nombreuses données expérimentales structurales et d'affinité sont disponibles pour Tiam1, ce qui en fait un très bon système test.

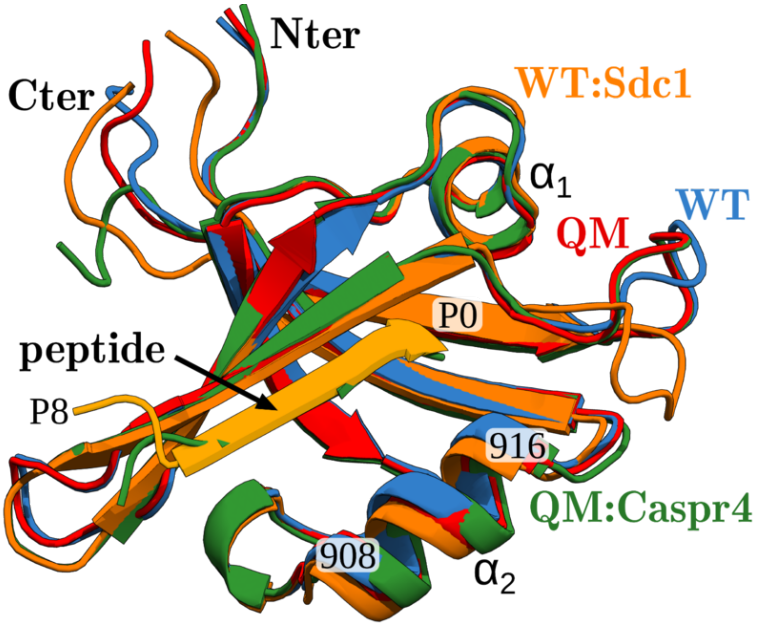


Figure 6.1 – Structures cristallographiques du domaine PDZ de Tiam1. Les structures des formes apo de Tiam1 WT (bleu) et QM (rouge) sont superposées aux structures des complexes WT:Sdc1 (orange) et QM:Caspr4 (vert).

6.2.1 Modèles structuraux

Sept structures cristallographiques du domaine PDZ de Tiam1 sont disponibles dans la PDB. On trouve la forme apo de Tiam1 (code PDB : 3KZD), les complexes Tiam1:Sdc1 (4GVD) et Tiam1:consensus (3KZE), correspondant à un peptide consensus déterminé à partir d'une bibliothèque combinatoire (voir partie 6.2.2). La structure d'un quadruple mutant (QM) de Tiam1 a également été résolue dans sa forme apo (4NXP) et liée aux peptides Caspr4 et Neurexine (4NXQ, 4NXR). Toutes ces structures présentent une résolution comprise entre 1,3 et 2,3 Å. Leur superposition (figure 6.1) montre que la liaison du peptide modifie très peu la conformation de la protéine. Les valeurs de RMSD entre les atomes du squelette situés au niveau de l'interface protéine-peptide (19 résidus à moins de 5 Å du peptide) sont ainsi comprises entre 0,5 et 1,5 Å. Le QM est modifié au niveau de quatre positions proches du peptide (911, 912, 915 et 920). La principale différence entre sa structure et la forme sauvage se situe au niveau de l'hélice α_2 qui est légèrement pivotée dans le cas du QM. Ce déplacement augmente la taille de la poche S_0 et lui permet d'accueillir des résidus plus grands à la position P_0 du peptide.

Figure 6.2 – Bibliothèque combinatoire des types reconnus par Tiam1 au niveau des cinq positions C-terminales du peptide. Chaque acide aminé est représenté par son code à une lettre dont la taille est proportionnelle à son occurrence. Les données sont issues de l'étude effectuée par Shepherd *et al.* [2011].

6.2.2 Affinités expérimentales

Pour optimiser les paramètres des modèles PB(GB)/LIE, des affinités expérimentales sont nécessaires. Le domaine PDZ de Tiam1 est un bon candidat puisque nous possédons 44 mesures d'affinité. 17 correspondent à des complexes dans lesquels la protéine Tiam1 présente une, deux ou quatre des mutations ponctuelles L911M, K912E, L915F et L920V, à l'interface protéine-peptide. Le peptide Sdc1 se lie à Tiam1 avec un K_d de 26,9 μM, ce qui représente une énergie libre de liaison de -6,70 kcal/mol. Si l'on exclut les deux systèmes présentant les affinités les plus faibles (2400 et 1600 μM), tous les complexes ont une affinité comprise entre 10,8 et 453 μM soit une plage d'énergies libres de 2,2 kcal/mol. Le complexe présentant la plus forte affinité est L920V:Caspr4, pour une énergie libre de liaison relative, de -0,54 kcal/mol par rapport au complexe WT:Sdc1. Trois peptides présentent une très faible affinité pour Tiam1 avec des valeurs de $\Delta\Delta G_b$ de 1,67 (Sdc2), 1,59 (Sdc4) et 1,56 kcal/mol (Sdc1-A0M). Pour six autres peptides non liants, la sensibilité de la mesure ne permet de donner qu'une borne inférieure à l'affinité mesurée de 250 μM, soit un $\Delta\Delta G_b$ de 1,32 kcal/mol ou plus. Toutes les autres valeurs d'énergies libres sont comprises entre -0,54 et 1,33 kcal/mol.

Une bibliothèque combinatoire des types reconnus aux cinq positions C-terminales du peptide est également disponible (Shepherd *et al.* [2011]). Le logo associé à cette bibliothèque (figure 6.2) révèle les types préférentiellement reconnus par Tiam1 à ces positions. Il montre que les positions P_0 et P_{-2} sont les plus spécifiques.

Le jeu d'énergies libres de liaison expérimentales a également été augmenté par des simulations alchimiques rigoureuses pour deux complexes, Tiam1:Sdc1-A0V et Tiam1:Sdc1-A0mA. Le calcul des affinités relatives de ces deux variants est décrite dans le chapitre 7.

6.3 Modélisation des complexes et simulations

6.3.1 Modélisation des complexes

51 complexes Tiam1:peptide ont été modélisés (tableaux 6.1 et 6.2). Les structures cristallographiques du domaine PDZ de Tiam1 lié aux peptides Syndecan1 (Sdc1), Caspr4, Neurexine (Neu) et au peptide consensus sont utilisées comme points de départ. Afin d'être le plus parcimonieux possible et limiter le nombre de modifications, la structure la plus proche du variant à modéliser est choisie. Ainsi, les dix complexes impliquant le peptide Caspr4 et son variant F0A ont été modélisés à partir de la structure du complexe QM:Caspr4. Les deux complexes liés à la Neurexine ont été modélisés à partir du complexe QM:Neu. Tous les autres complexes ont été modélisés à partir du complexe Tiam1:Sdc1.

Le positionnement des chaines latérales modifiées au niveau du peptide et/ou de la protéine est effectué avec le programme SCWRL4 (Krivov *et al.* [2009]). Les chaines latérales à moins de 4 Å des sites de mutations sont également autorisées à changer de rotamère. Dans la structure du complexe QM:Caspr4, l'extrémité N-terminale du peptide est désordonnée, la structure des deux premiers résidus n'est donc pas résolue. Les atomes manquants sont ajoutés en se basant sur la structure du complexe WT:Sdc1 puis leur position est ajustée à l'aide d'une étape de minimisation suivie d'une courte dynamique sous contraintes. Le reste du peptide Caspr4 a également une conformation différente de celle de Sdc1 (figure 6.3). Les 10 complexes impliquant Caspr4 sont construits à partir de ce modèle. L'extrémité N-terminale de chaque peptide est neutralisée par un groupement acétyle.

6.3.2 Modélisation de la 2-méthylalanine

La 2-méthylalanine (mAla ou Aib) est un acide aminé non présent dans les protéines et retrouvé dans certains polypeptides antibiotiques produits par les champignons. Il correspond à une alanine pour laquelle le H_α est remplacé par un méthyle. Il est capable d'augmenter la résistance des peptides aux protéases (Yamaguchi *et al.* [2003]), les rendant plus stables *in vivo*. Afin de déterminer si cette approche est applicable à Tiam1, des mutants des peptides Sdc1 et Caspr4 avec une mAla à la position P_0 ont étés modélisés.

Les paramètres de la mAla pour le champ de force Amber n'étant pas disponibles, une étape de paramétrisation a été nécessaire. Le modèle tridimensionnel de la mAla a été construit ma-

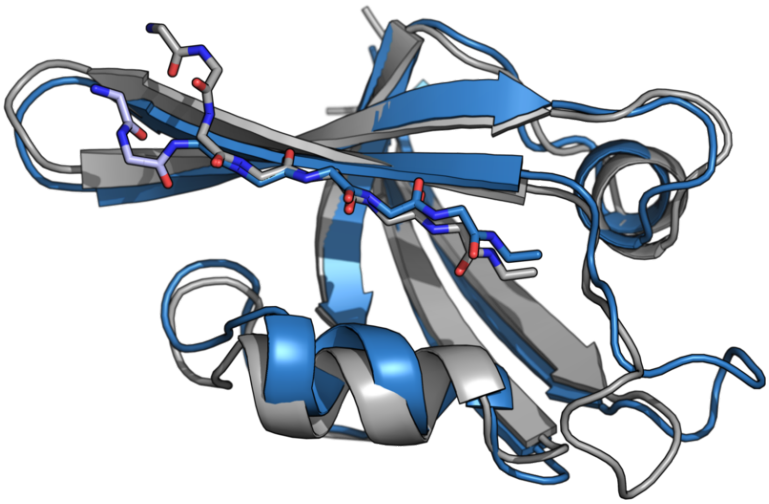


Figure 6.3 – Superposition des structures des complexes Tiam1:Sdc1 et QM:Caspr4. Pour chaque complexe, le squelette du peptide est représenté en bâtons. Tiam1:Sdc1 et QM:Caspr4 sont respectivement en gris et en bleu. Les atomes du peptide Caspr4 reconstruits sont en bleu clair.

nuellement à partir d'un résidu Ala acétylé en remplaçant le H_α par un groupement méthyle. Le modèle ainsi obtenu a ensuite été minimisé à l'aide du programme Gaussian 9 puis les charges partielles des différents atomes ont été déterminées par l'approche de Merz-Singh-Kollman (Singh & Kollman [1984]). Le modèle utilisé est celui de Hartree-Fock (Roothaan [1951]) avec la base 6-31G(d). Les charges partielles obtenues sont finalement légèrement ajustées à la main pour respecter la symétrie de la molécule. Ce niveau de description quantique est couramment utilisé pour étendre le champ de force ff99SB.

6.3.3 Simulations de dynamique moléculaire

Pour chaque complexe, une simulation de dynamique moléculaire en solvant explicite est effectuée. Les modèles produits par SCWRL4 sont préparés à l'aide du serveur Charmm GUI (Brooks *et al.* [2009] ; Sunhwan *et al.* [2008]). Les complexes sont immergés dans une boîte d'eau TIP3P (Jorgensen *et al.* [1983]) puis neutralisés à l'aide de quelques ions sodium. L'état de protonation des histidines est déterminé par inspection visuelle des structures et par le programme PropKa (Olsson *et al.* [2011]). Les systèmes sont ensuite minimisés pendant 1000 pas par la méthode du gradient conjugué en contraignant les atomes lourds puis en relâchant progressivement les contraintes. Une phase d'équilibration de 500 ps est ensuite effectuée en augmentant progressivement le pas d'intégration et la température et en supprimant progressivement les contraintes appliquées au squelette de la protéine. Les simulations sont effectuées

à température et pression constantes (300 K et 1 bar) en utilisant le thermostat et le barostat de Nosé-Hoover (Nosé [1984] ; Hoover [1985]). Les interactions électrostatiques sont traitées par la méthode du *Particle Mesh Ewald* ou PME (Darden *et al.* [1993]). Le champ de force Amber ff99SB a été utilisé (Cornell *et al.* [1996]). Des simulations de 40 à 100 ns ont été produites à l'aide du logiciel NAMD 2.12 (Phillips *et al.* [2005]), la longueur de chaque simulation dépendant du degré de convergence des composantes énergétiques utilisées par les modèles. Seuls les complexes Tiam1:Sdc1 et QM:Caspr4 sont prolongés jusqu'à 500 ns.

Lors des simulations, l'extrémité N-terminale du peptide est maintenue liée à Tiam1 par une contrainte non-invasive. L'énergie de contrainte est nulle tant que le peptide est à moins de 3 Å de la protéine. Au-delà de cette limite, une contrainte semi-harmonique est appliquée avec une constante de force de 3 kcal/mol/Å^2. Cette contrainte a été appliquée en raison du détachement occasionnel de l'extrémité N-terminale au cours de simulations tests, pouvant entrainer des fluctuations énergétiques importantes, impossibles à échantillonner correctement en 100 ns.

Pour 16 des 51 complexes, malgré les contraintes appliquées au niveau de l'extrémité N-terminale, la structure s'écarte de la structure cristallographique, notamment au niveau de l'hélice α_2. Le modèle PB/LIE mono-trajectoire que nous allons utiliser n'est pas capable de décrire l'énergie libre associée aux changements conformationnels pour des raisons que nous détaillerons plus loin. Pour ces complexes, des contraintes semi-harmoniques supplémentaires sont donc ajoutées pour maintenir les structures dans une conformation proche de la structure cristallographique. L'énergie de contrainte est prise en compte dans les calculs d'énergies libres de liaison. Ce terme est inférieur à 0,30 kcal/mol dans la plupart des cas, excepté pour les complexes L911M:Caspr4 et Tiam1:YAAGRKHF pour lesquels cette valeur est respectivement de 0,39 et 0,66 kcal/mol.

Un temps de simulation de 100 ns semble suffisant comme en attestent les valeurs de RMSD par rapport aux structures initiales (figure 6.4). De plus, les structures des 51 modèles restent très proches les unes des autres au cours des simulations puisque les valeurs RMSD calculées sur le squelette entre les structures moyennes de chaque modèle sont comprises entre 0,6 et 2,1 Å. L'interface protéine-peptide reste également très proche de la structure cristallographique avec des RMSD compris entre 0,7 et 1,5 Å.

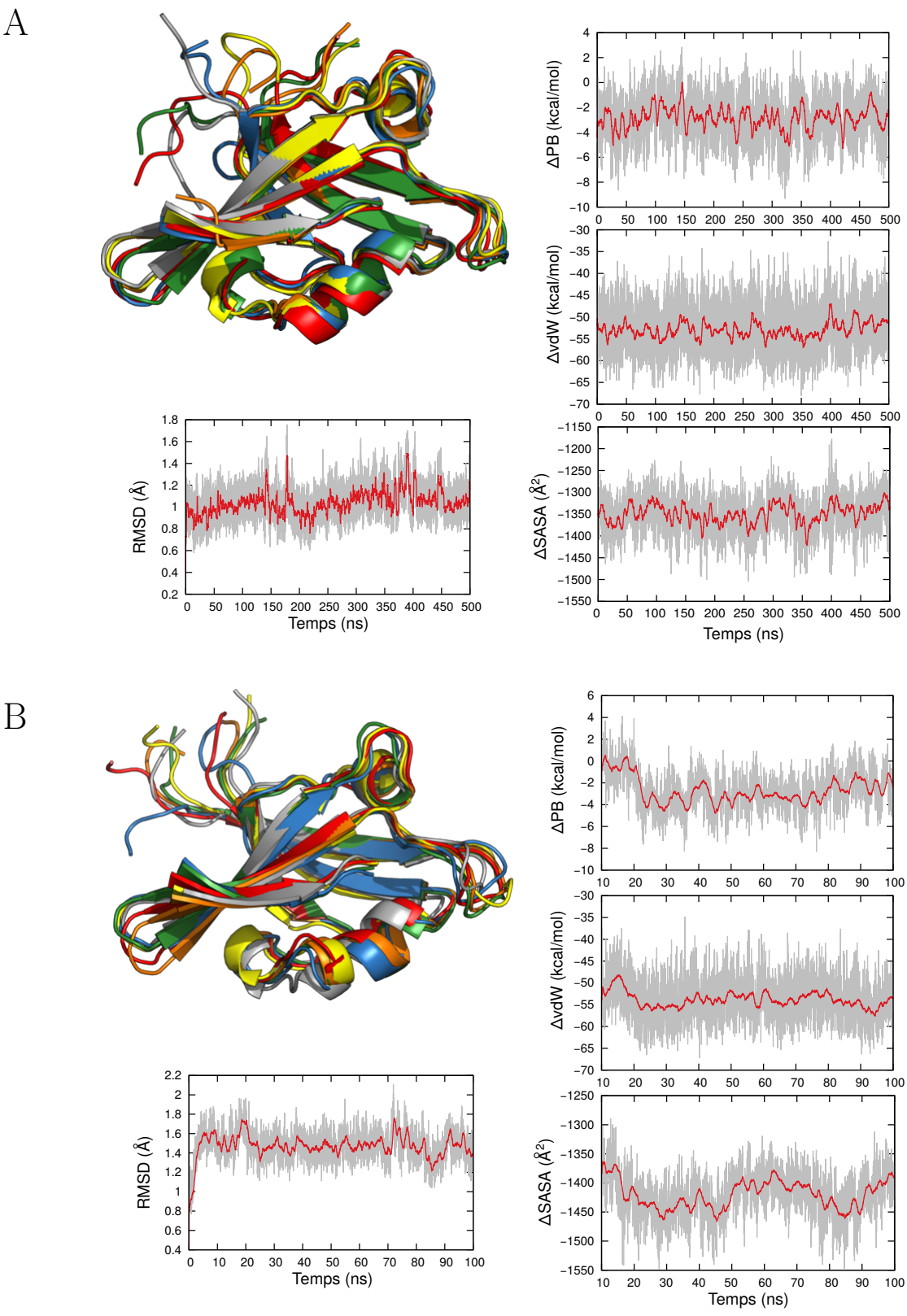


Figure 6.4 – Stabilité des structures et des composantes énergétiques au cours des simulations de dynamique moléculaire. Les complexes Tiam1:Sdc1 et QM:Sdc1 sont respectivement représentés en A et B. Pour chaque système, quelques structures issues de la dynamique moléculaire sont superposées. Le RMSD a été calculé par rapport à la structure initiale. Les termes énergétiques ont été extraits de la dynamique en prenant une conformation toutes les 20 ps.

6.3.4 Extraction des termes énergétiques

À partir des trajectoires de dynamique moléculaire, des conformations sont extraites toutes les 20 ps et utilisées pour calculer les différents termes énergétiques pris en compte dans les modèles. Les 10 premières nanosecondes sont systématiquement exclues pour permettre aux structures de se relaxer. Entre 1500 et 4500 conformations par simulation sont donc utilisées.

Nous utilisons ici la méthode de calcul mono-trajectoire, ce qui signifie que la simulation du complexe est utilisée pour décrire aussi bien l'état associé que dissocié. Pour cela, les composantes énergétiques sont successivement calculées en prenant en compte les atomes du complexe, du peptide et de la protéine. Le terme de van der Waals est directement extrait des énergies du champ de force. La contribution électrostatique ΔG_{elec} est estimée soit à l'aide d'un terme PB, soit d'un terme GB. Pour une conformation donnée, les molécules du solvant sont retirées et l'énergie libre électrostatique est calculée à partir d'un continuum électrostatique. La protéine et son ligand ont une constante diélectrique ϵ_P et le solvant une constante diélectrique ϵ_W. Lorsque le PB est utilisé, la valeur du potentiel électrostatique est calculée en résolvant numériquement l'équation de Poisson-Boltzmann en utilisant une grille cubique et la méthode des différences finies implémentée dans Charmm (Im *et al.* [1998]). La grille est composée de 181 plans dans chaque direction séparés de 0,8 Å. Les charges utilisées sont celles du champ de force Amber ff99SB. Le potentiel aux limites extérieures de la grille est approximé par le potentiel de Debye-Hückel produit par ces charges. Pour chaque structure un second calcul est effectué avec une maille plus fine (0,4 Å), en utilisant le potentiel de la grille issue du premier calcul comme condition aux bords. Cette méthode appelée *focusing* permet d'obtenir une meilleure estimation du potentiel électrostatique (Gilson *et al.* [1988]). La concentration en ions est fixée à 100 mM tandis que les constantes diélectriques du solvant (ϵ_W) et du soluté (ϵ_P) sont respectivement de 80 et 8. Pour délimiter les deux régions du système, nous utilisons ici un jeu de rayons atomiques optimisés pour le PB avec les champs de force Amber (Swanson *et al.* [2005]).

Pour les calculs GB, une version modifiée du programme Xplor a été utilisée (Brünger [1992]). Cette version implémente la méthode GB^{HCT} développée par Hawkins *et al.* [1995] qui est très similaire au variant utilisé dans Amber (Onufriev *et al.* [2002]). Le modèle a été paramétré pour être utilisé avec les charges de Amber (Lopes *et al.* [2007]).

L'erreur statistique des énergies libres calculées est obtenue en divisant chaque trajectoire en N partitions de 5 ns. L'incertitude est ensuite estimée de la manière suivante :

$$\sigma(\Delta G) = \sqrt{var(\Delta G)/(N-1)} \tag{6.2}$$

avec $var(\Delta G)$ la variance des N ΔG. Pour estimer l'incertitude de la valeur des $\Delta\Delta G_b$, la variance du complexe d'intérêt est ajoutée à celle de la référence. Toutes les incertitudes obtenues sont comprises entre 0,1 et 0,2 kcal/mol ce qui semble indiquer que les simulations sont suffisamment longues malgré les oscillations parfois observées pour certaines composantes (figure 6.4B).

6.4 Optimisation du modèle PB/LIE

Nous souhaitons optimiser les paramètres d'un modèle PB/LIE (équation 5.24) pour des complexes Tiam1:peptide. Plutôt que d'estimer l'énergie libre de liaison absolue, nous calculons l'énergie libre de liaison relative par rapport au complexe Tiam1:Sdc1. La qualité du jeu d'apprentissage impactant les performances du modèle produit, nous avons porté une attention particulière au choix des complexes utilisés pour l'optimisation des coefficients.

6.4.1 Jeu de données utilisé lors de l'optimisation

Des analyses préliminaires ont montré que les peptides ayant une affinité trop faible ($>$ 1,6 kcal/mol) étaient mal décrits par le modèle car, malgré les contraintes appliquées, ces peptides ont tendance à se détacher partiellement au cours des simulations. Le modèle mono-trajectoire suppose que la liaison du peptide n'entraine pas de changements structuraux importants. Ces peptides ne peuvent donc pas être évalués par le modèle et sont exclus du jeu d'apprentissage.

Les dix complexes mettant en jeu le peptide Caspr4 présentent une différence structurale systématique au niveau de la partie N-terminale du peptide par rapport aux autres complexes. Leurs structures ont été modélisées à partir du complexe QM:Caspr4 pour lequel la partie N-terminale du peptide était désordonnée. La différence de conformation entraine une erreur systématique lors des calculs d'énergies libres. Pour ces complexes, nous avons donc fait le choix de prendre comme référence non plus le complexe WT:Sdc1 mais le complexe QM:Caspr4. En

pratique, une correction est appliquée en ajoutant aux $\Delta\Delta G_b$ la valeur expérimentale de QM:Caspr4 et en soustrayant sa valeur calculée :

$$\Delta\Delta G_{\text{calc}}(X)' = \Delta\Delta G_{\text{calc}}(X) + [\Delta\Delta G_{\text{expt}}(\text{QM} : \text{Caspr4}) - \Delta\Delta G_{\text{calc}}(\text{QM} : \text{Caspr4})] \quad (6.3)$$

où X correspond au complexe impliquant Caspr4 (ou son variant F0A).

L'exclusion des peptides non-liants et du complexe QM:Caspr4 fait chuter à 38 les valeurs de $\Delta\Delta G_b$ disponibles. Les peptides impliquant la Neurexine sont également exclus de l'optimisation ce qui amène à 35 le nombre de valeurs de $\Delta\Delta G_b$ utilisées. La liste des 35 systèmes est présentée dans le tableau 6.1. Les peptides exclus du jeu d'apprentissage sont présentés dans le tableau 6.2.

6.4.2 Performances du modèle

Le modèle d'énergie libre a trois paramètres ajustables α, β et γ pondérant respectivement les termes PB, vdW et SA. Ces coefficients ont été choisis pour minimiser l'écart quadratique moyen entre les 35 valeurs expérimentales et calculées. Les coefficients optimaux sont $\alpha = 0{,}25$, $\beta = 0{,}020$ et $\gamma = -4$ cal/mol/Å^2 et sont comparables aux valeurs utilisées dans d'autres modèles LIE (Brandsdal *et al.* [2003] ; Carlsson *et al.* [2006]).

Les valeurs expérimentales sont comparées aux valeurs calculées en figure 6.5. Les paramètres et les erreurs statistiques du modèle sont présentées dans le tableau 6.3. Les valeurs moyennes des $\Delta\Delta G_b$ expérimentaux et calculés sont respectivement de 0,70 et 0,40 kcal/mol ce qui indique que le modèle surestime l'affinité des peptides de 0,30 kcal/mol en moyenne. Cette tendance est visible pour les mutants de Caspr4 et les peptides de faible affinité. L'écart quadratique moyen est de 0,55 kcal/mol tandis que l'erreur absolue moyenne est de 0,43 kcal/mol. Le coefficient de corrélation de Spearman est de 0,64. Les plus grandes erreurs observées sont de 1,31, 1,13 et 1,09 kcal/mol, dont deux variants de Caspr4. Lorsque l'on ne prend en compte que les variants de Caspr4, l'erreur moyenne absolue est de 0,59 kcal/mol. Nous avons confronté ces performances à un modèle Nul qui suppose que tous les peptides se lient à Tiam1 avec la même affinité (l'affinité moyenne). Ce modèle donne des erreurs proches du modèle PB/-LIE avec une erreur quadratique moyenne de 0,52 kcal/mol et une erreur absolue moyenne de 0,44 kcal/mol mais un coefficient de corrélation de zéro. Les plus grandes erreurs avec ce

Tableau 6.1 – Complexes Tiam1:peptide et énergies libres utilisés pour l'optimisation du modèle PB/LIE. Les énergies libres expérimentales et calculées, leur déviation et les différentes composantes énergétiques (PB, vdW et SA) sont présentées. Les énergies sont en kcal/mol et le terme SA en Å^2. Sauf quand cela est précisé, la protéine Tiam1 est dans sa forme sauvage.

complex	exp.	comp.		err.	PB	VdW	SA	rest.[a]	corr.[b]
Sdc1	0,00	0,00	(0,1)	0,00	0,00	0,00	0,00	0,00	0,00
Sdc1.A0F	0,43	0,16	(0,1)	-0,27	0,19	-4,21	-47,95	0,00	0,00
Sdc1.E4K	0,81	0,98	(0,2)	0,17	2,62	-0,17	-12,94	0,28	0,00
Sdc1.E4L	0,56	0,49	(0,2)	-0,07	1,95	-1,30	-8,08	0,00	0,00
Sdc1.E3D,Y1T	0,87	0,41	(0,1)	-0,46	1,67	2,08	12,05	0,00	0,00
Sdc1.E3T,Y1K	1,33	0,38	(0,1)	-0,95	1,49	3,49	15,64	0,00	0,00
Sdc1.F2I	0,80	0,26	(0,1)	-0,54	0,38	-0,44	-42,27	0,00	0,00
Sdc1.A0mA	0,04[c]	0,62	(0,1)	0,58	2,06	2,46	-13,51	0,00	0,00
Sdc3	0,13	0,21	(0,1)	0,08	0,38	1,38	-23,06	0,00	0,00
consensus	0,84	0,48	(0,1)	-0,36	2,72	3,35	67,97	0,00	0,00
YAAEKYWA	0,72	0,36	(0,1)	-0,36	2,10	4,73	64,29	0,00	0,00
YAAKAFRF	1,17	1,35	(0,1)	0,18	4,70	5,02	53,73	0,29	0,00
YAAYRYRA	1,32	1,26	(0,1)	-0,06	4,07	1,66	-18,18	0,14	0,00
YAARKFAK	1,32	1,30	(0,1)	-0,02	3,81	7,13	5,44	0,23	0,00
YAAKRTYV	1,32	1,14	(0,1)	-0,18	3,72	8,01	49,20	0,25	0,00
YAAGRKHF	1,32	1,52	(0,2)	0,20	3,49	2,11	14,15	0,66	0,00
YAALIHKF	1,32	0,98	(0,1)	-0,34	2,70	1,16	-2,87	0,27	0,00
YAAQKHFH	1,32	0,92	(0,2)	-0,40	2,59	-1,74	-34,81	0,17	0,00
QM:CADM1	0,87	1,43	(0,2)	0,56	5,74	4,77	24,95	0,00	0,00
L911M:Sdc1	0,15	0,32	(0,2)	0,17	0,62	-1,03	-45,29	0,00	0,00
K912E:Sdc1	0,97	0,58	(0,2)	-0,39	2,27	-0,99	-8,66	0,00	0,00
L911M,K912E:Sdc1	1,21	0,54	(0,1)	-0,67	1,71	-1,52	-35,42	0,00	0,00
L915F:Sdc1	0,65	0,05	(0,2)	-0,60	0,04	-0,98	-13,78	0,00	0,00
L920V:Sdc1	0,31	-0,07	(0,2)	-0,38	-1,10	0,17	-48,56	0,01	0,00
L915F,L920V:Sdc1	1,32	0,19	(0,2)	-1,13	-0,24	1,03	-27,04	0,12	0,00
QM:Sdc1	0,89	0,27	(0,1)	-0,62	-0,02	-0,83	-71,82	0,00	0,00
QM:Sdc1.A0F	0,22	0,20	(0,2)	-0,02	0,26	1,50	-26,96	0,00	0,00
QM:Caspr4	-0,23	-0,23	(0,1)	0,00	3,35	1,86	10,98	0,00	-1,06
WT:Caspr4	-0,21	-0,37	(0,1)	-0,16	2,04	2,77	32,77	0,26	-1,06
WT:Caspr4.F0A	0,52	-0,57	(0,1)	-1,09	2,90	3,61	77,81	0,00	-1,06
L911M:Caspr4	-0,39	0,00	(0,1)	0,39	2,81	3,49	24,68	0,39	-1,06
K912E:Caspr4	0,46	0,09	(0,1)	-0,37	2,70	-0,44	-50,13	0,28	-1,06
L911M,K912E:Caspr4	0,04	-0,01	(0,1)	-0,05	3,41	2,57	22,74	0,24	-1,06
L915F:Caspr4	0,48	-0,45	(0,1)	-0,93	2,33	0,78	-2,68	0,00	-1,06
L920V:Caspr4	-0,54	-0,51	(0,1)	0,03	2,29	-0,28	4,82	0,00	-1,06
L915F,L920V:Caspr4	0,62	-0,37	(0,1)	-0,99	2,49	-0,07	-17,92	0,00	-1,06
QM:Caspr4.F0A	1,09	-0,22	(0,1)	-1,31	4,59	3,78	95,58	0,00	-1,06

[a]Énergie de contrainte. [b]Correction de l'énergie libre (Eq. 6.3). [c]Valeur obtenue par simulation alchimique.

Tableau 6.2 – Complexes Tiam1:peptide et énergies libres de liaison non utilisés dans l'optimisation du modèle PB/LIE. Les énergies libres et les composantes énergétiques sont indiquées comme dans le tableau 6.1. Les énergies sont en kcal/mol, le terme SA en Å^2.

complex	exp.	comp.	err.	PB	VdW	SA	rest.[a]	corr.[b]
Sdc1.A0M	1,56	-0,05	-1,61	-0,51	-3,27	-36,08	0,00	0,00
Sdc1.A0V	1,90[c]	0,80	-1,10	1,30	-2,40	-81,36	0,20	0,00
Sdc2	1,67	0,37	-1,30	2,10	4,32	61,59	0,00	0,00
Sdc4	1,59	0,50	-1,09	1,69	4,46	59,45	0,23	0,00
QM :Neu	0,32	0,20	-0,12	0,70	3,98	12,96	0,00	0,00
L911M,K912E :Neu	1,25	0,29	-0,96	0,40	1,36	-39,92	0,00	0,00
L915F,L920V :Neu	1,08	0,17	-0,91	0,25	2,90	-11,94	0,00	0,00
Sdc1.A0Q	NA	0,18	NA	1,06	-2,72	7,76	0,00	0,00
Sdc1.F2C	NA	0,44	NA	1,15	0,81	-33,84	0,00	0,00
Sdc1.F2M	NA	0,01	NA	0,06	0,87	5,57	0,00	0,00
Sdc1.F2T	NA	0,17	NA	-0,05	-0,13	-45,20	0,00	0,00
Sdc1.F2V	NA	0,06	NA	-0,46	-1,09	-48,83	0,00	0,00
Sdc1.F2Y	NA	-0,04	NA	-0,57	-1,63	-33,98	0,00	0,00
Sdc1.E4L	NA	0,49	NA	1,95	-1,30	-8,08	0,00	0,00
Caspr4.F0mA	NA	0,09	NA	4,55	2,25	8,80	0,00	-1,06

[a]Énergie de contrainte. [b]Correction de l'énergie libre (Eq. 6.3). [c]Valeur obtenue par simulation alchimique.

Tableau 6.3 – Paramètres optimaux pour les différents modèles d'énergie libre et stastiques associées. Les paramètres ont été optimisés sur le jeu de 35 peptides en prenant comme références les complexes Tiam1:Sdc1 et QM:Caspr4.

α, β, γ	termes énergétiques	nom du modèle	rmsd	mue	R	$\mathrm{Err}_{\max}$	$\langle$ Err $\rangle$
NA	NA	Nul	0,52	0,44	0,00	1,1	0,00
0,25, 0,020, -4	PB+vdW+SA	PB/LIE	0,55	0,43	0,64	1,2	-0,30
0,26, 0,000, -2	PB+SA	PB/LIE	0,55	0,44	0,64	1,1	-0,30
0,14, 0,014, -5	GB+vdW+SA	GB/LIE	0,66	0,55	0,56	1,3	-0,43
0,20, 0,130, 0.05	GB+vdW+LK	GBLK	0,69	0,59	0,54	1,3	-0,48

Les énergies sont en kcal/mol ; SA en Å^2. $\mathrm{Err}_{\max}$ correpond à la moyenne des trois erreurs absolues les plus grandes.

modèle sont de 1,24, 1,09 et 0,91 kcal/mol et l'erreur absolue moyenne pour les variants de Caspr4 est de 0,56 kcal/mol.

Afin de déterminer la sensibilité du modèle au jeu de données, une validation croisée a été effectuée. Les 35 peptides sont séparés aléatoirement en 8 groupes de tailles égales. Ces groupes sont ensuite tour à tour exclus lors de l'optimisation des coefficients et utilisés comme jeu de validation. Les résultats sont présentés dans le tableau 6.4. La valeur des coefficients semble peu sensible aux peptides pris en compte même si le terme vdW subit de fortes fluctuations

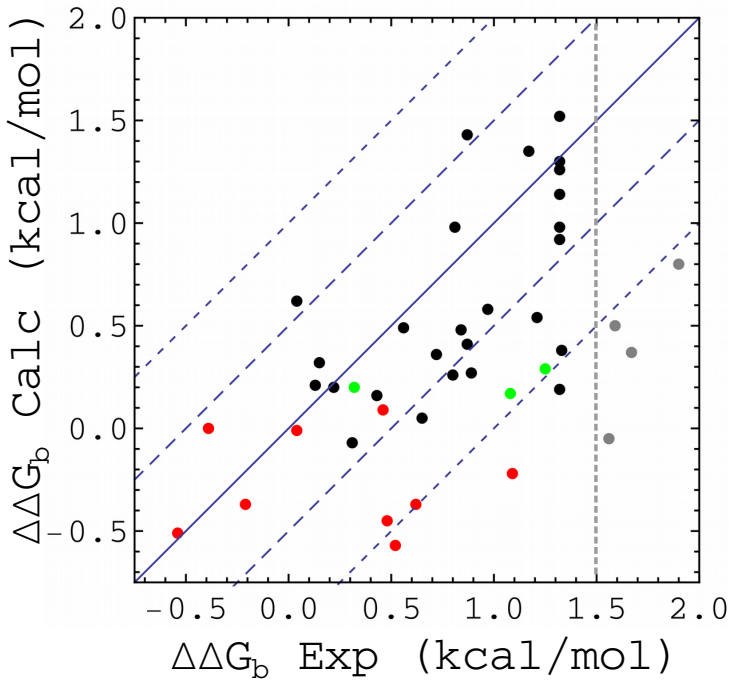


Figure 6.5 – Comparaison des énergies libres de liaison expérimentales et calculées à l'aide du modèle MMPBSA. Les termes PB, vdW et SA sont respectivement pondérés par les coefficients $\alpha = 0{,}25$, $\beta = 0{,}020$ et $\gamma = -4$ cal/mol/Å^2 . Rouge : complexes produits à partir de la structure QM:Caspr4 (4NXQ) ; Vert : complexes produits à partir de la structure QM:Neu (4NXR). Gris : complexes présentant une affinité faible pour Tiam1 (non pris en compte dans le modèle) ; Noir : tous les autres complexes.

et est absent de certains modèles. Les performances obtenues pour les différents modèles sur les échantillons d'apprentissage sont sensiblement les mêmes que celles du modèle complet et sont également très proches des résultats obtenus sur les échantillons de validation. Il ne semble donc pas y avoir de surapprentissage. Nous avons également appliqué le modèle à trois variants de la Neurexine qui n'ont pas été pris en compte lors de l'optimisation. L'erreur absolue moyenne obtenue est de 0,66 contre 0,44 kcal/mol avec le modèle Nul.

Un jeu de 100000 modèles aléatoires a été produit pour lesquels les valeurs expérimentales ont été permutées (Huang *et al.* [2006]). Pour chaque modèle, les coefficients α, β et γ ont été ajustés pour minimiser l'écart quadratique moyen entre les valeurs calculées et les valeurs expérimentales permutées. Les résultats sont présentés en figure 6.6. Parmi les modèles aléatoires produits, 99,94% donnent des erreurs plus importante et/ou un coefficient de corrélation plus faible.

Tableau 6.4 – Résultats de la validation croisée de modèle MMPBSA. Le jeu de données a été séparé en huit groupes de tailles similaires. Les groupes sont tour à tour exclus lors de l'optimisation des coefficients et utilisés comme jeu de validation. Les énergies sont en kcal/mol.

	Modèle			Apprentissage				Validation		
Groupes	α	β	γ^a	RMSD	MUE	Emax	R	RMSD	MUE	Emax
Tous	0,25	0,02	-4	0,56	0,44	1,18	0,65	-	-	-
1	0,21	0,07	-7	0,50	0,40	1,03	0,69	0,86	0,73	1,52
2	0,27	0,01	-5	0,55	0,43	1,16	0,63	0,62	0,52	1,19
3	0,26	0,00	-3	0,59	0,47	1,17	0,57	0,33	0,26	0,57
4	0,25	0,01	-3	0,53	0,41	1,15	0,65	0,76	0,69	0,97
5	0,24	0,03	-4	0,59	0,48	1,16	0,60	0,22	0,12	0,44
6	0,29	0,00	-3	0,57	0,44	1,16	0,63	0,54	0,50	0,72
7	0,26	0,01	-3	0,54	0,43	1,10	0,67	0,67	0,53	1,17
8	0,25	0,05	-6	0,58	0,46	1,21	0,75	0,42	0,37	0,68
Moyenne	0,25	0,02	-4	0,56	0,44	1,14	0,65	0,55	0,47	0,91
σ	0,02	0,03	2	0,03	0,03	0,06	0,06	0,22	0,21	0,37

^{a} : valeur en cal/mol/Å^2

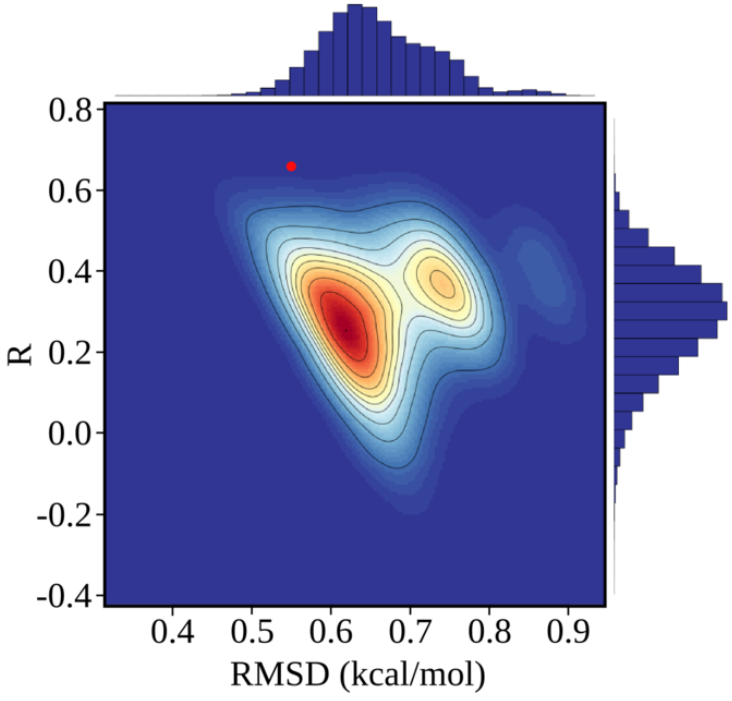


Figure 6.6 – Distribution des RMSD et des coefficients de corrélation des modèles aléatoires. La densité est indiquée par le dégradé du bleu au rouge. Les lignes correspondent aux déciles. Le modèle PB/LIE est représenté par le point rouge. Des histogrammes des valeurs de R et du RMSD obtenues pour les modèles aléatoires sont représentés sur le contour du graphique.

6.5 Modèles d'énergie libre alternatifs

D'autres modèles d'énergie libre ont également été testés. Ils diffèrent du PB/LIE, soit par les termes énergétiques utilisés, soit par le nombre de trajectoires utilisées pour modéliser les états associés et dissociés.

6.5.1 Traitements alternatifs des termes électrostatiques et apolaires

6.5.1.1 Traitement des interactions de van der Waals

Dans le modèle PB/LIE, la contribution du terme de van der Waals à l'énergie libre de liaison est faible. Un second modèle omettant le terme van der Waals a également été obtenu et donne des performances similaires au modèle à trois termes (tableau 6.3). Le coefficient associé au terme SA est négatif ($\gamma = -2$ cal/mol/Å^2), ce qui signifie que la surface enfouie lors de la liaison est pénalisée. En l'absence du terme de van der Waals, le terme PB est donc le seul à favoriser la liaison.

Jusqu'à présent, le terme de van der Waals prenait en compte uniquement les interactions protéine-peptide. Les interactions apolaires entre les solutés et le solvant sont traitées implicitement par les termes de vdW soluté-soluté et de SA. Les faibles erreurs obtenues avec le modèle PB/LIE semblent valider cette approximation. Nous avons cependant souhaité comparer les termes apolaires du modèle PB/LIE (vdW + SA) à l'énergie de van der Waals totale des solutés car cette dernière est parfois utilisée dans les modèles LIE. Cette comparaison a été effectuée sur dix peptides pour lesquels nous possédons des trajectoires pour les états associés et dissociés. La figure 6.7 montre le changement de l'énergie de van der Waals du peptide lors de la liaison et la compare au terme apolaire du modèle PB/LIE, $\beta\Delta E_{vdw} + \gamma\Delta A$. Ces deux quantités présentent un coefficient de corrélation de 0,9. L'utilisation des termes de vdW soluté-soluté et de SA permet donc bien d'approximer les interactions de vdW entre les solutés et le solvant.

6.5.1.2 Estimation de l'énergie libre de solvatation par un terme GB

Un modèle alternatif décrivant les interactions électrostatiques par un terme GB a été optimisé. Pour rappel, le GB est une approximation de l'équation de Poisson-Boltzmann qui

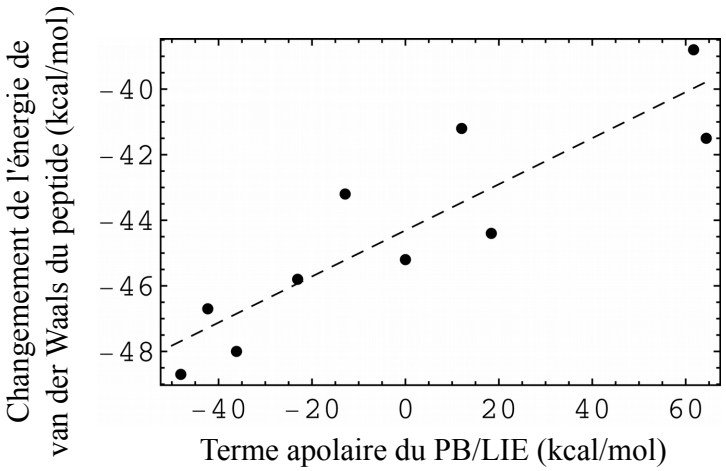


Figure 6.7 – Approximation du changement de l'énergie de van der Waals du peptide lors de la liaison (en incluant les interactions avec le solvant) par le terme apolaire du PB/LIE. Chaque point correspond à un peptide pour lesquels des trajectoires des états associés et dissociés étaient disponibles. Tous les peptides sont liés à la forme sauvage de Tiam1. La ligne en pointillé correspond à une régression linéaire (pente=0,07), indiquée pour plus de clarté.

calcule l'énergie électrostatique de solvatation plus rapidement. La version de GB utilisée ici est appelée GB^{HCT} (Hawkins *et al.* [1995]) et a été implémentée dans le logiciel Xplor (Brünger [1992]). La constante diélectrique du soluté est de 8. Les performances de ce modèle sont indiquées dans le tableau 6.3. Le modèle GB/LIE présente des erreurs légèrement plus importantes que celles obtenues à l'aide du PB avec une erreur absolue moyenne de 0,55 kcal/mol, un RMSD de 0,66 kcal/mol et un coefficient de corrélation de 0,56. Ce modèle tend également à surestimer l'affinité des complexes de 0,43 kcal/mol en moyenne. Le coefficient du terme électrostatique α est légèrement plus faible (0,14), le coefficient du terme de van der Waals est plus élevé (0,14) et le coefficient du terme surfacique est similaire (-5 cal/mol). Malgré des performances légèrement en retrait, les résultats du modèle GB/LIE restent très proches de ceux du modèle PB/LIE avec notamment un RMSD mutuel de 0,29 kcal/mol et un coefficient de corrélation de 0,87 (figure 6.8A).

6.5.1.3 Modélisation des interactions apolaires par une densité d'énergie

Les modèles PB/LIE et GB/LIE utilisent les termes de van der Waals et surfaciques pour capturer les interactions apolaires entre les solutés et le solvant. Le terme SA a été remplacé dans un modèle alternatif par un terme de densité d'énergie gaussienne proposé par Lazaridis

et Karplus (Lazaridis & Karplus [1999]) :

$$\begin{aligned} \Delta G_{LK} &= \sum_i G_i \\ G_i &= G_i^{\text{ref}} - \sum_{j \neq i} \int_{V_j} g_i(r_{ij}) dV \\ &= G_i^{\text{ref}} - \sum_{j \neq i} g_i(r_{ij}) V_j \end{aligned} \tag{6.4}$$

où la somme est calculée sur tous les atomes i du soluté et V_j correspond au volume de l'atome j. Chaque contribution reflète le transfert de l'atome i d'un état entièrement solvaté vers une conformation partiellement enfouie. L'énergie libre d'un atome i entièrement solvaté est donnée par une énergie de référence G_i^{ref} empirique. Le même atome, au sein du soluté, est écranté par les autres atomes du soluté qui réduisent son énergie de solvatation. Cette réduction est décrite par l'intégrale d'une densité d'énergie sur le volume des atomes environnants du soluté. La densité d'énergie a une forme gaussienne :

$$g_i(r_{ij}) = \frac{G_i^{\text{free}}}{2\pi^{3/2}\lambda_i r_{ij}^2} e^{-(r_{ij}-R_i)^2/\lambda_i^2} \tag{6.5}$$

où r_{ij} est la distance interatomique, R_i est le rayon de l'atome i et λ_i une distance de corrélation. Le paramètre G_i^{free} est tel que, lorsque i est entièrement enfoui, l'énergie libre de solvatation est égale à zéro. L'énergie libre totale a la forme :

$$\Delta G_{LK} = \sum_i G_i^{\text{ref}} - \sum_{i,j \neq i} g_i(r_{ij}) V_j \tag{6.6}$$

Les paramètres utilisés pour le terme LK, ont préalablement été optimisées à partir de mutations de stabilité et de structures (Michael *et al.* [2017]). Le terme LK est calculé à partir des conformations issues des trajectoires puis utilisé à la place du terme SA.

En combinant ce modèle avec un terme GB, les performances obtenues sur le modèle des 35 peptides sont légèrement moins bonnes que celles du modèle GB/LIE avec une erreur quadratique moyenne de 0,69 kcal/mol, une erreur moyenne absolue de 0,55 kcal/mol et un coefficient de corrélation de 0,54 (tableau 6.3). Le modèle reste proche du modèle PB/LIE avec un RMSD mutuel de 0,33 kcal/mol et un coefficient de corrélation de 0,84 (figure 6.8B) entre les deux modèles.

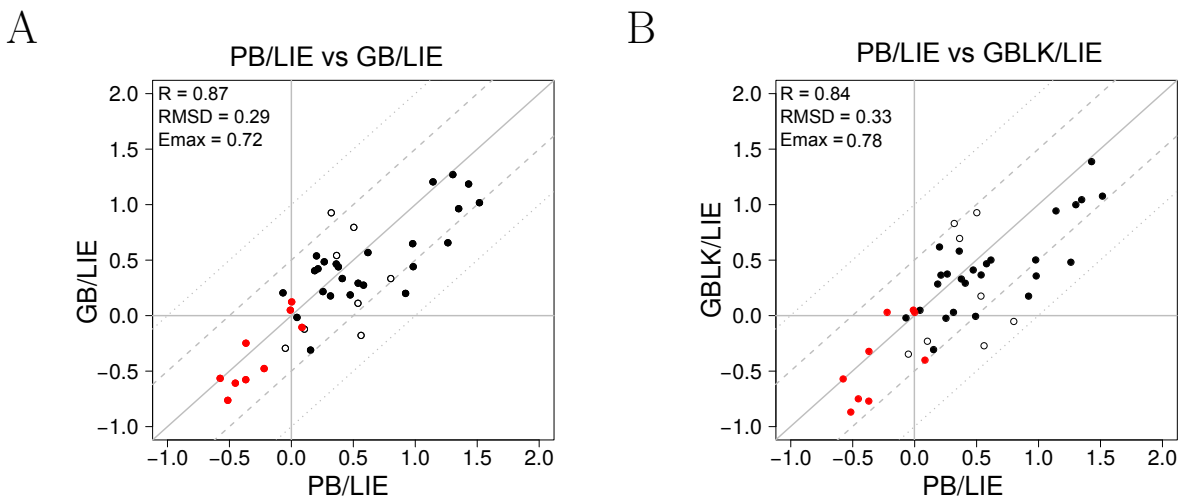


Figure 6.8 – Comparaison des modèles PB/LIE, GB/LIE et GBLK. Les valeurs de $\Delta\Delta G_b$ ont été estimées par les différents modèles dont les coefficients sont indiqués dans le tableau 6.3. Les points rouges correspondent aux complexes modélisés à partir du complexe QM:Caspr4, les cercles correspondent aux peptides non liants et les points noirs à tous les autres systèmes.

6.5.2 Modification de l'échantillonnage

6.5.2.1 Protocole à trois trajectoires

Les modèles utilisés jusqu'à présent sont basés sur l'approche mono-trajectoire. Cette approche n'est applicable que lorsque la liaison du ligand n'entraine pas de changements conformationnels importants au niveau des solutés. La comparaison des structures cristallographiques des formes apo et holo du domaine PDZ de Tiam1 semble indiquer que cette approche est applicable à ce système. Pour les complexes Tiam1:Sdc1 et QM:Caspr4 nous avons toutefois appliqué un modèle à trois trajectoires afin de savoir si les résultats obtenus sont comparables au modèle mono-trajectoire. Pour améliorer l'échantillonnage, des simulations plus longues ont été effectuées. Ainsi, l'état dissocié de chaque peptide a été simulé pendant 400 ns, les deux complexes pendant 500 ns et l'état dissocié de la protéine pendant 1 μs. Les composantes énergétiques sont extraites des trajectoires de la même manière que pour le protocole mono-trajectoire, excepté pour le terme PB qui nécessite un traitement particulier. En effet, lors du calcul PB, chaque charge atomique est répartie sur plusieurs nœuds du maillage. Lorsque le potentiel électrostatique est calculé, les différentes parties d'une même charge vont donc interagir entre elles. Cet artefact s'annule si le potentiel des états associé et dissocié a été calculé en utilisant la même grille et la même conformation, ce qui est vrai dans le calcul mono-trajectoire. Au contraire, dans le protocole à trois trajectoires, les confor-

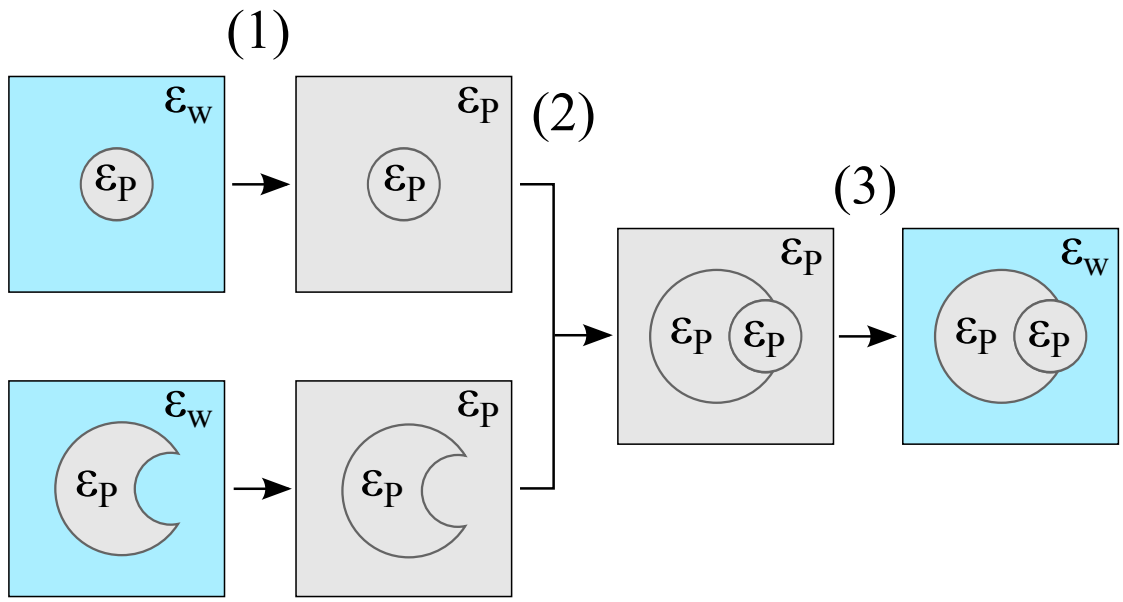


Figure 6.9 – Estimation du terme d'énergie électrostatique dans le protocole à trois trajectoires. Les étapes effectuées sont les suivantes : (1) l'énergie libre associée au transfert des deux partenaires du solvant (ϵ_W) vers un milieu ayant la même constante diélectrique que la protéine (ϵ_P) est calculée ; (2) l'énergie libre associée à la liaison des deux partenaires dans un milieu diélectrique ϵ_P est calculée à partir du terme de Coulomb ; (3) l'énergie libre associée à la restauration de la constante électrique du solvant lorsque les deux partenaires sont liés est calculée.

mations étant différentes pour les différents systèmes (complexe, protéine, ligand), il n'est pas possible de corriger directement cet artefact. Il est alors nécessaire de calculer la composante électrostatique en trois étapes : (1) on évalue l'énergie libre associée au transfert des deux partenaires du solvant ($\epsilon_W = 80$) vers un milieu ayant la même constante diélectrique que la protéine ($\epsilon_P = 8$) ; (2) on évalue l'énergie libre associée à la liaison des deux partenaires dans un milieu diélectrique $\epsilon_P = 8$; (3) on évalue l'énergie libre associée à la restauration de la constante électrique du solvant lorsque les deux partenaires sont liés (figure 6.9). La contribution des étapes (1) et (3) est obtenue en résolvant l'équation de PB à l'aide de Charmm. La contribution de l'étape (2) est déterminée en calculant la différence du terme énergétique de Coulomb entre l'état associé et l'état dissocié, divisée par ϵ_P.

La différence d'énergie de liaison entre les deux systèmes est de 1,04 kcal/mol, ce qui est très proche de la valeur de 0,99 kcal/mol obtenue avec le modèle mono-trajectoire (tableau 6.5). Bien que les résultats soient très proches, la contribution des composantes n'est pas la même entre les deux approches. Dans les deux cas la composante électrostatique est responsable de la majeure partie de la différence d'énergie libre mais présente une valeur deux fois plus

Tableau 6.5 – Comparaison des énergies libres de liaison obtenues par les approches mono et 3-trajectoires en utilisant le modèle PB/LIE. Les complexes, protéines et peptides ont respectivement été simulés pendant 500, 1000 et 400 ns. Les coefficients du modèle PB/LIE sont α=0.25, β=0.02 et γ=-4 kca/mol/Å^2. Les énergies sont en kcal/mol.

Protocole	Complexe	PB	VdW	SA	ΔG_b	$\Delta\Delta G_b$
1-traj	Tiam1 :Sdc1	-2.84	-58.67	-1349.29	3.51	0.99
	QM :Caspr4	0.25	-58.29	-1399.72	4.50	
3-traj	Tiam1 :Sdc1	-4.22	-47.61	-1216.89	2.86	1.04
	QM :Caspr4	1.62	-41.94	-1084.73	3.90	

importante dans le protocole à trois trajectoires. La composante van der Waals est quinze fois plus élevée dans le protocole à trois trajectoires mais sa contribution finale reste faible en raison du faible coefficient β. Les valeurs plus importantes des composantes PB et van der Waals dans le protocole à trois trajectoires sont compensées par le terme surfacique également plus élevé mais favorisant cette fois le complexe QM:Caspr4 quand les deux autres sont favorables à Tiam1:Sdc1.

Par ailleurs, la prolongation des simulations des complexes à 500 ns montre que dans l'approche mono-trajectoire, les valeurs obtenues sont seulement 0,16 kcal/mol au-dessus de la valeur calculée à partir des trajectoires de 100 ns, en accord avec l'erreur statistique estimée précédemment. Cette faible différence semble confirmer que les simulations de 100 ns sont suffisantes pour obtenir des résultats convergés.

6.5.2.2 Prise en compte de la réorganisation du peptide

Avec l'approche mono-trajectoire, la conformation dissociée du peptide est extraite du complexe. Les composantes de l'énergie libre sont alors calculées à partir de la forme étendue du peptide puisque ce dernier forme un feuillet β avec Tiam1. Les changements structuraux du peptide dissociés sont modélisés de manière implicite par la constante diélectrique ϵ_P. Cette approche suppose que les changements conformationnels sont faibles (Simonson [2013] ; Swanson *et al.* [2004]). Les simulations du peptide dissocié ayant montré qu'il était très flexible, un modèle plus détaillé peut être nécessaire. Ce modèle sépare le processus de liaison en deux étapes (figure 6.10) : (1) le peptide adopte une conformation étendue capable de se lier à la protéine ; (2) le peptide se lie à la protéine. L'énergie libre de la première étape notée ΔG_I, ou énergie libre conformationnelle, peut être estimée à partir de la fraction étendue observée

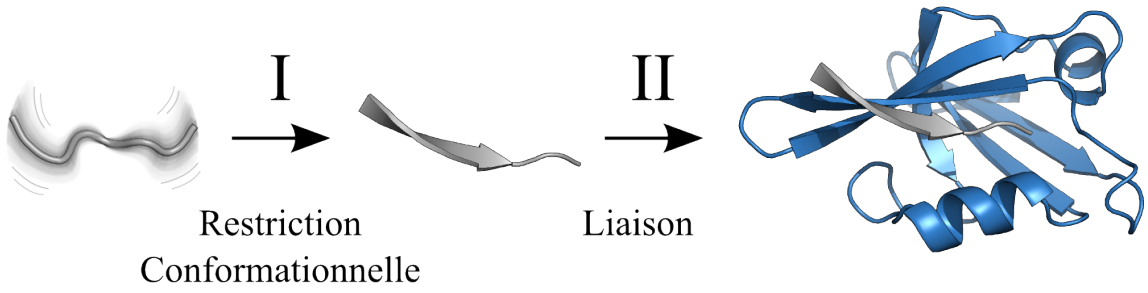


Figure 6.10 – Représention schématique du processus de liaison en deux étapes. Le peptide adopte la conformation qu'il occupe dans l'état lié (étape 1) puis se lie à la protéine (étape 2).

dans les simulations du peptide en solution. La seconde contribution peut être estimée à avec un modèle PB/LIE ou GB/LIE. Nous qualifions ce modèle de modèle à deux trajectoires.

Estimation de l'énergie libre de repliement L'énergie libre conformationnelle ΔG_I est calculée à partir de longues simulations du peptide en solution. La fraction du temps que le peptide passe dans son état étendu est déterminée en se basant sur les valeurs des angles ϕ et ψ des cinq résidus C-terminaux. Si tous ces angles sont dans la région du diagramme de Ramachandran correspondant aux brins β ($\psi \geq 60°$ ou $\leq$ -150° et $\phi \leq$ -30°), le peptide est considéré comme étendu, capable de lier la protéine. Pour calculer la différence d'énergie libre conformationnelle entre deux peptides i et j, les fractions étendues f_i et f_j sont déterminées puis la différence d'énergie libre est estimée de la manière suivante :

$$\Delta\Delta G_I(i,j) = -k_B T \log f_i/f_j \tag{6.7}$$

où k_B correspond à la constante de Boltzmann et T à la température (300 K dans notre cas).

Performances du modèle à deux trajectoires L'énergie libre conformationnelle a été estimée pour 12 peptides, dont 11 sont capables de se lier à la forme sauvage de Tiam1. Le dernier (Sdc1-F2R) présente une faible énergie libre de liaison et est considéré comme non liant. Pour chaque peptide, deux simulations indépendantes de 100 ou 200 ns sont effectuées. Les énergies libres obtenues sont comprises entre 0,0 kcal/mol pour le peptide de référence Sdc1 et 1,2 kcal/mol pour les peptides Sdc3 et YAAEKYWA, avec une énergie libre moyenne de 0,7 kcal/mol (tableau 6.6). Les valeurs positives indiquent que les variants sont tous moins structurés que le peptide Sdc1. L'incertitude estimée en comparant les deux simulations est

Tableau 6.6 – Énergies libres de réorganisation des peptides. Les énergies libres ont été estimées à partir de simualtions des peptides en solution. La fraction repliée est déterminées à partir des valeurs des angles ϕ et ψ des 5 résidus C-terminaux.

peptide	% replié[b]	ΔG_I
[a]Sdc1	5,8/3,8	0,0 (0,2)
A0F	3,4/0,3	0,9 (0,8)
A0M	4,0/0,8	0,6 (0,6)
E4K	5,0/3,2	0,1 (0,3)
F2I	1,1/1,1	0,9 (0,1)
F2R	2,4/0,7	0,8 (0,5)
E3D,Y1T	4,6/0,6	0,6 (0,7)
Sdc2	2,0/0,8	0,8 (0,4)
Sdc3	0,7/0,6	1,2 (0,1)
[a]Caspr4	2,6/8,6	0,0 (0,5)
[a]Caspr4-F0A	5,7/1,6	0,3 (0,5)
[a]YAAEKYWA	8,4/0,1	1,2 (1,6)

ΔG_I (kcal/mol) correspond à l'étape I dans la figure 6.10. [a]Ces peptides ont été simulés pendant 2×200 ns, les autres pendant 2×100 ns. [b]Le peptide 'replié' présente une conformation étendue proche de celle observée dans la forme liée.

comprise entre ±0,1 et ±0,7 kcal/mol (±0,4 kcal/mol en moyenne) excepté pour le peptide YAAEKYWA qui, malgré les 400 ns de simulation, présente une incertitude de ±1,6 kcal/mol. Les incertitudes sont du même ordre de grandeur que les énergies libres calculées. Il est donc difficile d'extraire une information de ces simulations. Cela est principalement dû à la grande flexibilité du peptide qui explore rarement l'état étendu.

Les 12 peptides simulés sont impliqués dans 25 complexes de notre jeu de données. Nous avons donc optimisé les coefficients α, β, γ du modèle PB/LIE sur ce jeu de données restreint en prenant en compte ou non la contribution de ΔG_I. Cette contribution n'améliore pas les résultats puisque l'écart quadratique moyen passe de 0,60 à 0,68 kcal/mol et le coefficient de corrélation passe de 0,52 à 0,26. La méthode est très dépendante de la qualité de l'échantillonnage. Une autre méthode, plus couteuse, consisterait à contraindre le peptide dans sa forme étendue puis à relâcher progressivement les contraintes.

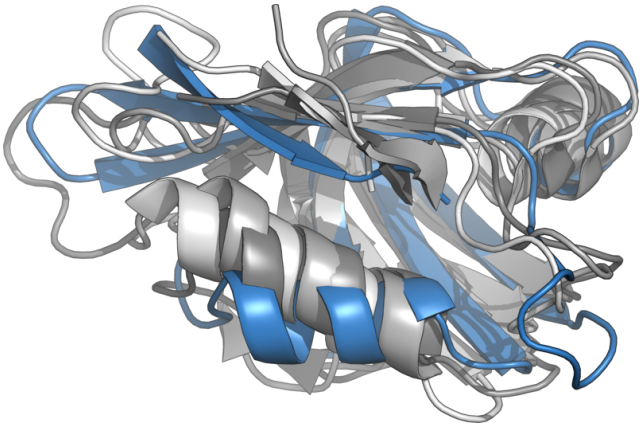


Figure 6.11 – Comparaison de la structure de Tiam1 avec celles des domaines PDZ liant un peptide possédant une valine en P_0. Trois des huits structures analysées sont représentées en gris (3JXT, 2OQS et 2AIN). La structure de Tiam1 liée à Sdc1 est représentée en bleu.

6.6 Analyse des structures et des énergies libres

6.6.1 Analyse des structures

Au regard des simulations effectuées, on peut émettre des hypothèses quant aux affinités observées.

Trois variants de la position P_0 du peptide Sdc1 ont été modélisés avec les résidus Val, Met et Phe. Sdc1-A0V n'a pas été caractérisé expérimentalement mais une affinité très faible a été prédite par transformation alchimique (FEP) : $\Delta\Delta G_b = 1{,}90$ kcal/mol. Ce résultat est conforté par l'absence de Val à position P_0 dans la librairie combinatoire (figure 6.2, Shepherd *et al.* [2011]). La mutation A0V n'entraine pas de changement structural important. Néanmoins, huit complexes PDZ:peptide issus de la PBD (IBE9, 3PDV, 4G69, 3QE1, 3JXT, 2OQS, 2AIN et 1VJ6) ayant une valine à la position P_0 présentent une orientation légèrement différente de l'hélice α_2 (figure 6.11). Cette différence pourrait augmenter le volume de la poche S_0 et favoriser ainsi la liaison de la valine. La capacité à lier la valine ne semble pas provenir des résidus formant la poche S_0 puisque ces derniers sont très similaires à ceux de Tiam1 dans les huit complexes étudiés. Le peptide Sdc1-A0M se lie également faiblement à Tiam1 ($\Delta\Delta G_b =$ 1,56 kcal/mol). La modélisation de ce variant est validée par les transformations alchimiques qui reproduisent l'affinité expérimentale (chapitre 7). Au cours de la simulation, la méthionine interagit avec L915 à l'extrémité C-terminale de l'hélice α_2, entrainant une déstructuration partielle de l'hélice. Cette observation ne semble pas en faveur d'une bonne affinité. Dans le

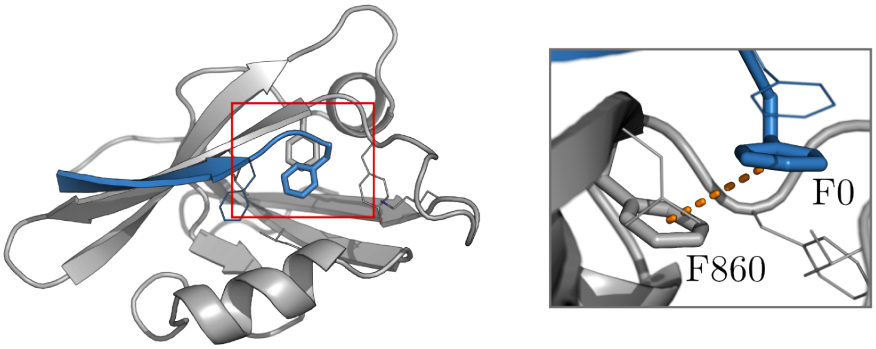


Figure 6.12 – Structure du complexe Tiam1:Sdc1-A0F. Le domaine PDZ est représenté en gris tandis que le peptide est en bleu. La figure de droite correspond à un agrandissement montrant l'interaction entre les résidus F860 et F0.

variant Sdc1-A0F, la chaine latérale de F0 perturbe également l'hélice α_2. Mais ce variant se lie à Tiam1 plus fortement que Sdc1-A0M ($\Delta\Delta G_b$ = 0,43 kcal/mol). La déstructuration de l'hélice α_2 pourrait donc être contrebalancée par l'interaction favorable entre les résidus F0 et F860 (π-*stacking*) qui est présente 50% du temps au cours de la simulation (figure 6.12).

Quelques variants de Sdc1 à la position P_{-2} ont également été étudiés. Le mutant Sdc1-F2I possède une affinité plus faible que Sdc1 avec $\Delta\Delta G_b$ = 0,80 kcal/mol. Au cours de la simulation, l'isoleucine conserve sa position dans la poche S_{-2} avec une orientation proche de la phénylalanine sauvage. Néanmoins, l'isoleucine présente une mobilité accrue et son interaction avec L911 dans l'hélice α_2 déforme légèrement cette dernière. Enfin, l'isoleucine pouvant explorer plus de rotamères que Phe, il est possible que sa perte d'entropie suite à la liaison du peptide soit plus importante. La liaison du variant Sdc1-F2R n'est pas détectable expérimentalement bien que l'Arg soit présente dans la bibliothèque combinatoire. Au cours de la simulation, l'arginine interagit avec le glutamate à la position P_{-4} réduisant l'interaction de ce résidu avec la protéine. En effet, le peptide sauvage interagit avec les résidus R871 et S908 respectivement 38% et 75% du temps. Ces populations sont de 3% (R871) et 56% (S908) dans le variant Sdc1-F2R. Cette observation est compatible avec l'idée d'une coopérativité négative ou positive entre les résidus du peptide. Cela pourrait expliquer pourquoi le mutant Sdc1-F2R ne se lie pas à Tiam1 alors que l'on retrouve quelques Arg à la position P_{-2} dans la bibliothèque combinatoire (figure 6.2).

Le mutant Sdc1-E4K a un $\Delta\Delta G_b$ de 0,81 kcal/mol qui semble principalement dû à la perte des interactions E4-R871 et E4-S908 observées dans la structure cristallographique sauvage. Le groupement amine repousse la lysine K912 située au niveau de α_2. La perte des interactions

E4-R871 et E4-S908 dans le cas du mutant Sdc1-E4L pourrait également expliquer la valeur du $\Delta\Delta G_b$ observée pour ce mutant (0,56 kcal/mol).

Finalement, le double mutant Sdc1-E3T,Y1K se lie faiblement à Tiam1 ($\Delta\Delta G_b$ = 1,33 kcal/mol). La chaine latérale de T3 interagit très peu avec la protéine alors que E3 présent dans le sauvage interagit 16% du temps avec le résidu N876. De plus, la lysine en position P_{-1} entraine une répulsion de R879 à 9 Å de sa position d'origine. Ces observations pourraient expliquer la faible affinité de ce mutant pour Tiam1.

6.6.2 Décomposition de l'énergie libre

Afin d'identifier les résidus jouant un rôle important dans la reconnaissance des peptides, la contribution des différents résidus à l'énergie libre de liaison a été calculée. Nous nous sommes intéressés aux résidus présents à l'interface protéine-peptide pour six complexes : Sdc1, Sdc1-A0M, Sdc1-A0F, Sdc1-A0mA, Caspr4 et Caspr4-F0A. Pour rappel, les séquences des peptides Sdc1 et Caspr4 sont TKQEEFYA et ENQKEYFF. Les résultats sont présentés dans le tableau 6.7. Parmi les résidus étudiés, certaines interactions sont identifiées comme étant importantes. Ainsi, la position P_0 semble favoriser les résidus de petite taille puisque l'alanine donne l'énergie libre la plus faible. On constate cependant que les interactions de van der Waals sont favorables aux résidus de grande taille (Met et Phe) en raison d'interactions avec les résidus hydrophobes de la poche et notamment L915. Les positions P_{-1} et P_{-3} sont très similaires entre les deux peptides, les variations observées semblent dans ce cas provenir de la flexibilité ces chaines latérales qui sont exposées au solvant. Les résidus à la position P_{-2} de Sdc1 et Caspr4 sont très proches et correspondent respectivement à une Tyr et une Phe. La contribution de cette position à l'énergie libre de liaison semble directement liée à la nature du résidu à la position P_0, les résidus de petite taille étant défavorables. Cela est majoritairement dû à la composante SA, plus élevée lorsqu'une Ala ou une mAla est présente en P_0.

La présence d'un Glu à la position P_{-4} des variants Sdc1 favorise la liaison du peptide. Cette contribution favorable est due aux interactions électrostatiques avec les résidus R871, K912 et, dans une moindre mesure, S908. Les peptides Sdc1 et Caspr4 possèdent tous les deux une Gln à la position P_{-5}, cependant leur contribution à l'énergie libre de liaison est très différente. Cet écart provient de la différence de conformation des squelettes des deux peptides. En effet, les deux chaines latérales ne sont pas orientées du même côté du brin β.

Tableau 6.7 – Décomposition des contributions à l'énergie libre de liaison des résidus de l'interface protéine-peptide, Les termes PB, vdW et SA sont pondérés par les coefficients du modèle PB/LIE,

Pos	P-7	P-6	P-5	P-4	P-3	P-2	P-1	P0	858	860	866	871	908	912	915
Composante PB															
Sdc1	-0,39	-0,05	-0,09	-1,19	-0,21	-0,67	0,01	0,01	0,31	0,07	-0,19	-0,40	-0,28	-0,11	0,31
A0F	-0,32	-0,11	-0,02	-1,06	-0,22	-0,75	-0,04	0,07	0,32	-0,02	-0,18	-0,20	-0,21	-0,13	0,07
A0M	-0,39	-0,07	-0,25	-0,96	-0,17	-0,70	-0,01	0,00	0,29	0,02	-0,24	-0,31	-0,17	-0,04	0,04
Sdc1 Dia	-0,37	-0,02	-0,12	-0,98	-0,13	-0,63	-0,01	-0,08	-0,02	-0,25	-0,20	-0,33	-0,22	-0,11	0,17
Caspr Dia	-0,21	-0,90	-0,10	-0,35	-0,19	-0,46	0,03	-0,09	-0,07	-0,15	0,01	-0,24	-0,02	0,05	0,14
Caspr	-0,31	-0,93	-0,11	-0,22	-0,19	-0,74	-0,05	0,08	-0,32	-0,02	0,09	-0,24	-0,32	0,07	0,13
Composante vdW															
Sdc1	-0,11	-0,23	-0,40	-0,06	-0,22	-0,64	-0,30	-0,21	-0,11	-0,17	-0,18	-0,14	0,00	-0,10	-0,26
A0F	-0,11	-0,23	-0,36	-0,05	-0,22	-0,56	-0,33	-0,78	-0,12	-0,23	-0,17	-0,08	-0,02	-0,11	-0,16
A0M	-0,12	-0,23	-0,31	-0,08	-0,22	-0,62	-0,31	-0,68	-0,12	-0,26	-0,17	-0,08	-0,02	-0,05	-0,19
Sdc1 Dia	-0,11	-0,19	-0,34	-0,08	-0,22	-0,63	-0,28	-0,34	-0,07	-0,12	-0,17	-0,11	-0,01	-0,09	-0,18
Caspr Dia	-0,20	-0,11	-0,19	-0,22	-0,20	-0,71	-0,32	-0,38	-0,07	-0,18	-0,18	-0,17	-0,08	-0,06	-0,17
Caspr	-0,18	-0,08	-0,19	-0,22	-0,20	-0,51	-0,21	-0,79	-0,11	-0,25	-0,18	-0,17	-0,03	-0,17	-0,18
Composante SA															
Sdc1	0,25	0,41	0,83	0,66	0,34	1,81	0,61	0,88	0,03	0,22	0,56	0,38	0,36	0,30	0,39
A0F	0,28	0,44	0,80	0,55	0,27	1,39	0,60	1,75	0,09	0,30	0,56	0,21	0,41	0,35	0,18
A0M	0,26	0,46	0,81	0,57	0,28	1,56	0,59	1,51	0,08	0,29	0,56	0,25	0,41	0,16	0,28
Sdc1 Dia	0,31	0,38	0,80	0,65	0,35	1,82	0,51	1,43	0,03	0,22	0,58	0,33	0,38	0,28	0,39
Caspr Dia	0,62	0,61	0,35	0,58	0,44	1,87	0,73	1,36	0,04	0,23	0,56	0,43	0,44	0,23	0,30
Caspr	0,53	0,52	0,35	0,51	0,41	1,49	0,47	1,73	0,07	0,30	0,58	0,48	0,40	0,42	0,20
Contribution au ΔG_b															
Sdc1	-0,25	0,14	0,35	-0,59	-0,08	0,51	0,31	0,68	0,23	0,12	0,18	-0,17	0,09	0,09	0,44
A0F	-0,15	0,10	0,42	-0,56	-0,17	0,07	0,24	1,04	0,29	0,05	0,21	-0,07	0,19	0,12	0,09
A0M	-0,24	0,15	0,25	-0,47	-0,11	0,24	0,27	0,83	0,26	0,06	0,14	-0,14	0,22	0,08	0,13
Sdc1 Dia	-0,17	0,16	0,33	-0,41	-0,00	0,56	0,22	1,00	-0,02	-0,15	0,20	-0,11	0,15	0,08	0,38
Caspr Dia	0,21	-0,40	0,07	0,01	0,05	0,70	0,43	0,90	0,04	-0,10	0,40	0,02	0,29	0,22	0,27
Caspr	0,04	-0,49	0,05	0,06	0,02	0,23	0,20	1,01	0,28	0,04	0,49	0,08	0,05	0,32	0,15

Cette orientation défavorable pour Caspr4 pourrait en partie expliquer la nécessité d'appliquer un terme de correction négatif dans le modèle PB/LIE. L'orientation de la chaine latérale de la position P_{-6} est également différente entre les peptides Sdc1 et Caspr4, ce qui favorise la liaison du peptide Caspr4 en raison d'interactions électrostatiques favorables avec le résidu R871. Enfin, la présence d'un glutamate à la position P_{-7} semble nuire à la liaison du peptide Caspr4 en raison d'interactions électrostatiques défavorables avec le résidu E866.

6.6.3 Prédiction de nouveaux variants

Le modèle PB/LIE a été optimisé dans le but de prédire l'affinité de nouveaux variants de Sdc1. Le modèle a donc été appliqué à de nouveaux variants de Sdc1 à la position P_{-2} (Cys, Met, Thr, Val et Tyr) (tableau 6.2). Dans le cas de F2C, l'énergie libre prédite est moins favorable que le peptide sauvage avec $\Delta\Delta G_b = 0{,}4$ kcal/mol. Les autres variants ont des affinités identiques ou légèrement plus faibles (0,2 kcal/mol au plus). Nous prédisons notamment que l'acide aminé non naturel mAla peut remplacer Phe à la position P_0 de Caspr4 avec une légère perte d'affinité de 0,1 kcal/mol par rapport au complexe WT:Sdc1, ou 0,3 kcal/mol par rapport

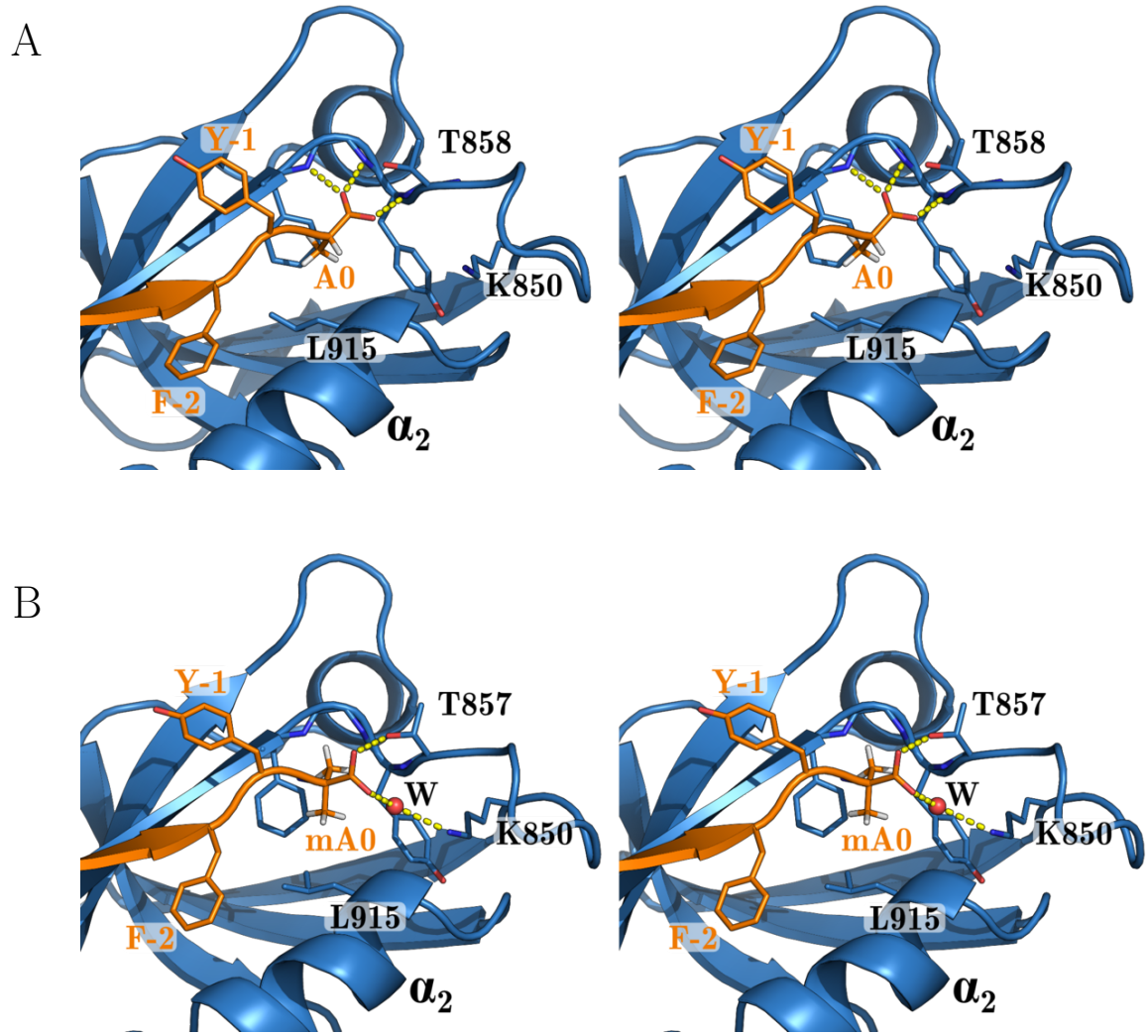


Figure 6.13 – Structures des complexes Tiam1:Sdc1 et Tiam1:Sdc1-A0mA. Les complexes Tiam1:Sdc1 (A) et Tiam1:Sdc1-A0mA (B) sont représentés en stéréo.

au complexe QM:Caspr4. Ces valeurs sont très proches des énergies libres de liaison calculées en utilisant la méthode plus rigoureuse de transformation alchimique. Ce type de mutation est intéressant car il pourrait rendre le peptide résistant aux protéases et donc augmenter sa durée de vie *in vivo* (Welch *et al.* [2007]). Dans le cas du complexe Tiam1:Sdc1-A0mA, l'extrémité C-terminale sort légèrement de la poche P_0. Cette conformation est stabilisée par une molécule d'eau formant une liaison hydrogène à la fois avec la chaine latérale de K850 et l'extrémité C-terminale du peptide (figure 6.13). Ce déplacement a été observé dans plusieurs simulations indépendantes et est réversible lorsque l'on mute la mAla vers une Ala.

6.7 Conclusions et discussion

Le but principal de cette étude était de paramétrer et tester une classe de modèles d'énergie libre pour des complexes PDZ:peptide. L'estimation de l'énergie libre par des approches computationnelles reste un défi majeur. Les approches par simulations alchimiques, qui ne présentent pas de paramètres ajustables, donnent de bons résultats mais sont très couteuses, notamment pour comparer des peptides très différents comme Sdc1 et Caspr4. Nous nous sommes donc intéressés à une classe de modèles semi-empiriques, moins couteuse et applicable à un plus grand nombre de systèmes. Ces modèles allient des simulations de dynamique moléculaire en solvant explicite à une fonction d'énergie libre simple. Différents variants existent et utilisent généralement un terme PB ou GB pour décrire les interactions électrostatiques du solvant, avec un terme de van der Waals et/ou surfacique. Les modèles peuvent également être complexifiés par l'ajout de termes supplémentaires permettant par exemple de décrire les liaisons hydrogène ou encore d'estimer l'entropie vibrationnelle.

L'étude de complexes protéine:peptide a l'avantage de s'appuyer sur des champs de force bien établis. La modélisation de ces complexes ne demande donc pas d'étape de paramétrisation. Ils présentent néanmoins certaines difficultés notamment en raison de la taille et la complexité de l'interface de liaison et de la flexibilité du peptide dissocié. Le modèle PB/LIE repose sur 35 affinités expérimentales mesurées sur des formes sauvages et mutées de Tiam1. Ces affinités s'étendent sur une petite plage (1,87 kcal/mol) ce qui représente une difficulté supplémentaire. Le modèle PB/LIE ainsi que ses variants GB/LIE et GBLK possèdent au plus trois paramètres ajustables. Nous avons également fait le choix d'appliquer l'approche mono-trajectoire afin de réduire les temps de calculs.

La modélisation des complexes est un point crucial pour la qualité des prédictions. Un mauvais modèle a de grandes chances de mener à une estimation erronée de l'énergie libre de liaison. Afin de s'assurer de la qualité de nos modèles, nous avons, dans un premier temps, testé leur stabilité par de longues simulations de dynamique moléculaire. Pour valider les modèles de deux complexes, des simulations alchimiques ont été effectuées et donnent des affinités très proches des mesures expérimentales. Le modèle mono-trajectoire étant incapable de décrire les changements conformationnels de grandes ampleurs, l'ajout de faibles contraintes est nécessaire. Cela est cependant applicable uniquement lorsque les contraintes n'ont qu'un faible impact sur l'énergie et la dynamique du système.

Le modèle PB/LIE présente une erreur quadratique moyenne de 0,55 kcal/mol, une erreur moyenne absolue de 0,43 kcal/mol et un coefficient de corrélation de Spearman de 0,65. Le modèle nul donne des erreurs moyennes similaires mais une corrélation de zéro. En supposant que l'erreur expérimentale moyenne est de $\delta_{exp} = 0{,}2$ kcal/mol, que les simulations de dynamique moléculaire introduisent une erreur aléatoire, $\delta_{MD} = 0{,}2$ kcal/mol, en notant σ_{tot} l'erreur quadratique moyenne et δ_{model} l'erreur introduite par le modèle d'énergie libre, on obtient $\sigma^2_{\mathrm{tot}} = 0.55^2 = \delta^2_{\mathrm{exp}} + \delta^2_{\mathrm{MD}} + \delta^2_{\mathrm{model}}$, soit $\delta_{\mathrm{model}} = 0.47$ kcal/mol. Cela correspond à environ deux fois l'incertitude expérimentale et à un quart de la plage des énergies libres expérimentales. Les modèles GB/LIE et GBLK donnent des résultats similaires au modèle PB/LIE. Le modèle à deux trajectoires, bien que théoriquement plus juste, n'améliore pas les résultats et augmente considérablement l'incertitude des valeurs calculées. Cela illustre la difficulté à échantillonner correctement l'espace conformationnel des molécules très flexibles. Lors d'études préliminaires nous avons également tenté d'estimer l'entropie vibrationnelle à partir des trajectoires de dynamique moléculaire en utilisant l'approximation quasi-harmonique. Cela a mené à des résultats non convergés et inexploitables.

Aucun des modèles développés n'est capable de prédire l'affinité des variants les moins affins (K_d de l'ordre du millimolaire). Il est probable que dans ces cas précis, la liaison du peptide nécessite des changements conformationnels trop importants pour être pris en compte dans nos modèles ou que le mode de liaison de ces peptides est différent de celui des peptides Sdc1 et Caspr4.

Le modèle PB/LIE n'a pas encore permis d'identifier de nouveaux peptides ayant une forte affinité pour le domaine PDZ de Tiam1. Il semble néanmoins possible d'insérer un acide aminé mAla à l'extrémité C-terminale des peptides Sdc1 et Caspr4 sans dégrader leur affinité. Cette modification pourrait rendre le peptide résistant aux protéases et ainsi augmenter sa stabilité *in vivo*. Les bonnes performances du modèle PB/LIE suggèrent qu'il pourra être utilisé pour prédire l'affinité de nouveaux variants mais également pour interpréter les résultats expérimentaux.

Chapitre 7

Calcul d'affinité par la méthode de transformation alchimique

Comme nous avons pu le voir dans le chapitre précédent, les approches de type LIE (*Linear Interaction Energy*) et MM/PBSA, sont largement utilisées pour déterminer *in silico* l'affinité entre un ligand et une protéine. Malgré les bons résultats qu'elles permettent d'obtenir, ces méthodes nécessitent certaines approximations, notamment au niveau de la description du solvant ou encore en négligeant la contribution du terme entropique (Gilson & Zhou [2007]). Cela rend leur utilisation difficilement applicable dans certains cas, en particulier si le système subit des changements conformationnels importants suite à la liaison du ligand. Une autre limitation de ces approches est qu'il n'existe a priori pas de jeu de coefficients applicables à tous les systèmes. Il est donc nécessaire de posséder initialement des valeurs expérimentales pour optimiser dans un premier temps les coefficients. Se pose alors la question de l'impact du choix des données expérimentales à la fois sur la qualité du modèle, mais également sur sa transférabilité à des systèmes proches.

D'autres méthodes, plus rigoureuses, basées sur la théorie thermodynamique des perturbations ont été développées (Landau & Lifshitz [1938] ; Beveridge & DiCapua [1989]). Cette théorie décrit la perturbation provoquée par le passage d'un système d'un état A vers un état B. On parle alors de perturbation de l'énergie libre (ou FEP pour *Free energy perturbation*) (Zwanzig [1954] ; Straatsma & McCammon [1992] ; Kollman [1993]). Ces approches sont beaucoup plus couteuses en temps de calcul par rapport aux modèles semi-empiriques mais n'ont pas de paramètres ajustables. Elles nécessitent toutefois d'avoir à sa disposition un champ de force capable de décrire de manière juste les interactions entre les atomes du système.

Dans ce chapitre nous nous intéresserons à une approche particulière et largement utilisée appelée transformation alchimique (Tembe & McCammon [1984]). Cette méthode a été de nombreuses fois utilisée pour calculer l'énergie libre de liaison absolue ou relative de complexes protéine:ligand (Woo & Roux [2005] ; Wan *et al.* [2005] ; Ioannidis *et al.* [2016] ; Aldeghi *et al.* [2016]). Des simulations alchimiques ont été effectuées pour 22 variants des peptides Sdc1 et Caspr4 et un variant de la protéine lié à ces deux peptides afin de calculer l'énergie libre de liaison relative. Parmi ces variants, 12 possèdent des valeurs d'affinité expérimentale et permettront de déterminer les performances de la méthodes mais également de valider les modèles structuraux. L'étude des autres variants permettra potentiellement d'identifier des peptides ayant une affinité accrue pour le domaine PDZ de Tiam1. Dans un second temps, la même approche sera utilisée pour calculer l'affinité de la forme phosphorylée du peptide Sdc1, révélant une limitation importante du champ de force ff99SB.

7.1 Méthodes

7.1.1 Modèles structuraux

Les structures cristallographiques des complexes Tiam1:Sdc1 (4GVD, chaines A et D) et du quadruple mutant QM:Caspr4 (4NXQ, chaines A et D) ont été utilisées pour modéliser 14 mutants ponctuels des peptides Sdc1 et Caspr4 ainsi qu'un variant de Tiam1, K912E, lié aux deux peptides (Liu *et al.* [2013, 2016]). La structure du complexe Tiam1:Sdc1 présente des résidus manquants au niveau des boucles flexibles β_1-β_2 et β_2-β_3. Ces deux boucles ont , dans un premier temps, été reconstruites à l'aide du programme MODELLER (Sali & Blundell [1993] ; Fiser *et al.* [2000]) en utilisant comme modèles la chaine B de la structure 4GVD et la structure apo de Tiam1 (3KZD) (Shepherd *et al.* [2010]). Les mutants ont ensuite été modélisés avec le programme SCWRL4 (Krivov *et al.* [2009]), les complexes Tiam1:Sdc1 (4GVD) et QM:Caspr4 servant respectivement à modéliser les mutants de Sdc1 et Caspr4. Ces mutants sont localisés au niveau des cinq positions C-terminales des peptides qui sont responsables de la spécificité de la reconnaissance des domaines PDZ pour leur ligand (Songyang *et al.* [1997] ; Stiffler *et al.* [2007] ; Tonikian *et al.* [2008]). Parmi les mutants modélisés, deux mutants doubles sont étudiés, Sdc1-E3D,Y1T et Sdc1-E3T,Y1K. Dans le cas du mutant Sdc1-E3D,Y1T les mutants simples Sdc1-E3D et Sdc1-Y1T sont également produits. La modélisation des mutants A0M et A0F

de Sdc1 entraine un changement de rotamère du résidu L915 qui se retrouve alors exposé au solvant, modifiant la poche de liaison S_0 (voir résultats). Pour ces deux mutants, un deuxième modèle structural est produit en contraignant l'orientation de la chaine latérale de L915 dans son orientation cristallographique. Afin d'identifier le modèle structural le plus plausible, ces deux modèles sont étudiés par la suite. Les peptides Sdc1 et Caspr4 ainsi que leurs variants ont également été modélisés dans leur état dissocié. Pour cela, les coordonnées des atomes des peptides sont extraites des complexes une fois les mutations introduites. La même opération est effectuée pour le variant K912E, mais en conservant cette fois les coordonnées de la protéine.

7.1.2 Simulations de dynamique moléculaire

Les complexes modélisés ont été dans un premier temps simulés par dynamique moléculaire. Ces simulations permettront de confirmer la stabilité des modèles produits. Elles serviront également de références pour valider l'espace conformationnel exploré au cours des transformations alchimiques. Les modèles produits sont préparés à l'aide du serveur Charmm GUI (Brooks *et al.* [2009] ; Sunhwan *et al.* [2008]). Les complexes sont immergés dans une boîte d'eau TIP3P (Jorgensen *et al.* [1983]) octaédrique puis neutralisés par quelques ions sodium. Les systèmes sont ensuite minimisés pendant 1000 pas par la méthode du gradient conjugué en contraignant les atomes lourds puis en relâchant progressivement les contraintes. Une phase d'équilibration de 500 ps est ensuite effectuée en augmentant progressivement le pas d'intégration ainsi que la température et en supprimant progressivement les contraintes appliquées sur le squelette de la proteine. Les simulations sont effectuées à température et pression constantes (300 K et 1 bar) en utilisant le thermostat et le barostat de Nosé-Hoover (Nosé [1984] ; Hoover [1985]). Les interactions électrostatiques sont traitées par la méthode du *Particle Mesh Ewald* ou PME (Darden *et al.* [1993]). Le champ de force Amber ff99SB a été utilisé (Cornell *et al.* [1996]). Des simulations de 40 à 500 ns ont ainsi été produites à l'aide du logiciel NAMD 2.12 (Phillips *et al.* [2005]).

7.1.3 Calculs d'énergie libre

7.1.3.1 Transformations alchimiques

Les énergies libres de liaison relatives des 16 variants de Sdc1, Caspr4 et Tiam1 ont été calculées par la méthode de transformation alchimique. Ces variants correspondent soit à des

mutants ponctuels, soit à des doubles mutants. Au total, 22 transformations alchimiques ont été effectuées. L'énergie libre de mutation associée aux transformations a été calculée au cours de simulations du complexe et du peptide libre. L'énergie libre relative de liaison, $\Delta\Delta G_b$, est calculée de la manière suivante :

$$\Delta\Delta G_b = \Delta G_{mut}^{complexe} - \Delta G_{mut}^{peptide} \tag{7.1}$$

Pour calculer l'énergie libre de mutation, nous avons utilisé le paradigme de la topologie double. Le résidu modifié lors de la transformation est ainsi composé de deux chaines latérales n'interagissant pas entre elles et positionnées sur un squelette commun au niveau du C_α. Les interactions de van der Waals et électrostatiques des deux chaines latérales avec le reste du système sont pondérées par un paramètre de couplage λ. Pour une valeur de λ donnée on a :

$$U(\lambda) = (1 - \lambda)U_A + \lambda U_B \tag{7.2}$$

où A et B correspondent aux deux mutants.

Les simulations alchimiques ont été effectuées à l'aide du logiciel NAMD 2.12 en utilisant la fonction `alch` (Liu *et al.* [2012]). La transformation est découpée en 11 valeurs de λ (0 ; 0,1 ; 0,2 ; ... ; 0,9 ; 1) que l'on appellera par la suite fenêtres. Afin d'assurer une insertion/délétion graduelle des chaines latérale, les interactions de type van der Waals des chaines latérales sont traitées par la méthode *soft-core* (Zacharias *et al.* [1994]). Pour assurer la stabilité des simulations, les interactions électrostatiques de la seconde chaine latérale ne sont activées qu'à partir de la fenêtre $\lambda = 0{,}4$. Cela permet d'éviter les interactions électrostatiques trop fortes entre les particules apparaissant et les particules existantes du système aux faibles valeurs de λ (Beutler *et al.* [1994]). L'énergie libre associée aux transformations a été estimée par la méthode du ratio d'acceptation de Bennett (BAR, Bennett [1976]) qui détermine le meilleur estimateur minimisant la variance de l'énergie libre (chapitre 5 partie 5.1.1.3).

7.1.3.2 Organisation des transformations

Pour produire les trajectoires aux différentes valeurs de λ deux stratégies ont été utilisées. La première, dite parallèle, consiste à effectuer les simulations des 11 fenêtres de manière indépendante pour des durées de 2 ns. À l'issue des simulations, une nouvelle série de simulations

est initiée en utilisant comme point de départ les conformations finales de la série précédente. La conformation utilisée peut soit être celle correspondant à la même valeur λ, soit à celle d'un λ adjacent ($\lambda \pm 0{,}1$). Cette dernière option améliore l'échantillonnage aux différentes fenêtres. Dix à quinze séries de transformations, soit 200-300 ns, sont nécessaires pour obtenir des résultats convergés pour les complexes ou la protéine seule. Les simulations des peptides isolés ne nécessitent que cinq séries pour converger, soit 110 ns. Pour initialiser la première série, les structures finales des dynamiques simples sont utilisées comme points de départ pour les fenêtres extrêmes (0 ; 0,1 ; 0,2 et 0,8 ; 0,9 ; 1). Pour démarrer les fenêtres intermédiaires une courte simulation à $\lambda = 0{,}25$ est effectuée pour que la seconde chaine latérale apparaisse progressivement et se positionne correctement. Ce protocole présente l'avantage d'être rapide car les fenêtres de la transformation étant indépendantes, elles peuvent être produites simultanément. En contre partie, dans le cas où le passage de $\lambda = 0$ à $\lambda = 1$ nécessite un changement conformationnel important, ce dernier peut être mal échantillonné ce qui aboutit à des résultats erronés.

Une autre approche dite séquentielle consiste à aller progressivement de $\lambda = 0$ à $\lambda = 1$. Pour cela, les fenêtres sont produites les unes après les autres en utilisant la conformation finale du λ précédent comme point de départ. Cette méthode est plus lente car les différentes fenêtres ne sont pas indépendantes. Pour que le système puisse se réorganiser à chaque fenêtre, des fenêtres de 10 ns sont produites. Afin d'étudier la réversibilité de la transformation, un aller-retour au moins est effectué ce qui représente au total 220 ns. Cette méthode a l'avantage de limiter les problèmes d'échantillonnage en retraçant explicitement le chemin de l'état A vers l'état B. Il est alors aisé de savoir si la transformation est réversible en comparant les structures obtenues pour les valeurs de λ=0 et 1 à celles des dynamiques simples. Cette seconde approche est privilégiée lors de l'insertion d'un résidu de grande taille demandant un réarrangement conformationnel de la protéine.

7.1.3.3 Estimation de l'erreur

Bootstrap L'erreur statistique associée aux calculs d'énergie libre est estimée par la méthode de rééchantillonnage du *bootstrap* (Jain *et al.* [1987]). Pour chaque fenêtre, 2000 conformations sont tirées au hasard parmi l'ensemble des conformations puis l'énergie libre est calculée à l'aide du BAR. La procédure est répétée 500 fois puis l'erreur est déterminée en calculant l'écart type entre les valeurs d'énergie libre obtenues.

Erreur de fermeture L'énergie libre de mutation entre deux systèmes est théoriquement indépendante du chemin de la transformation. Par exemple, pour trois systèmes A, B et C, il est possible de calculer l'énergie libre associée à la transformation de A vers B de deux façons différentes : (1) par le chemin direct entre A et B ; (2) en deux étapes par le chemin A vers C, puis C vers B, et en sommant ensuite l'énergie libre associée aux deux transformations. L'énergie libre calculée par ces deux approches devrait en théorie être la même. Dans le cas contraire, cela signifie qu'au moins l'un des deux chemins ne décrit pas de manière exacte la transformation. L'erreur de fermeture permet donc d'avoir une information sur la qualité de l'échantillonnage (Mobley & Klimovich [2012] ; Wang *et al.* [2013]).

7.1.4 Décalage du potentiel lié au PME

Les simulations de dynamique moléculaire sont généralement effectuées en appliquant des conditions périodiques aux limites de la boîte. Dans ce cas, le système est décrit comme une boîte aux limites finies entourées d'une grille infinie d'images périodiques. Cette approche constitue le seul moyen de simuler de manière formelle un système infini. Dans un tel système, une façon de calculer les interactions électrostatiques consiste à utiliser la sommation d'Ewald (de Leeuw *et al.* [1980]), ou son approximation, le PME (Darden *et al.* [1993]), qui prend en compte explicitement l'ensemble des particules de la boîte et de toutes ses images. Le système est alors décrit comme un ensemble de boîtes de simulation contenues dans un milieu uniforme (communément appelé *gellium*) de charge opposée à la charge du système. Ainsi, si une charge q est insérée dans le système, une charge de densité $-q/V$ (où V est le volume de la boîte) sera introduite de manière uniforme de sorte à neutraliser la charge du système. Cette correction décale artificiellement le potentiel électrostatique de la boîte de simulation (Yen-Lin *et al.* [2014]).

L'énergie libre associée à l'insertion d'une charge dépend explicitement du décalage du potentiel électrostatique qui dépend lui-même de la nature du soluté et de son volume par rapport au volume de la boîte (Yen-Lin *et al.* [2014]). La contribution du décalage doit donc être prise en compte pour les mutations ioniques puisque le volume du complexe est plus important que le volume du peptide isolé.

Pour estimer la contribution du PME à la différence d'énergie libre des mutations ioniques, le potentiel électrostatique au cours de simulations d'un peptide isolé et de deux complexes

Tiam1:peptide a été calculé. Le peptide Sdc1-F2R et les complexes Sdc1-F2I et Sdc1-F2R ont été simulés pendant 10 ns dans une boîte d'eau cubique dont la fraction d'eau est identique à celle de la boîte octaédrique. Les potentiels électrostatiques ont ensuite été moyennés sur les conformations issues de ces dynamiques à l'aide du module PMEPot de VMD (Aksimentiev & Schulten [2005] ; Humphrey *et al.* [1996]). Nous ferons par la suite référence aux systèmes neutre et chargé par P^0 et P^+ respectivement. À partir des résultats de PMEPot, le potentiel électrostatique, $\overline{\Phi}_W$, est moyenné sur la boîte cubique en prenant en compte uniquement la région à plus de 12 Å de la protéine. Pour chacun des systèmes, le décalage provoqué par le PME, $\delta\Phi$, dans la simulation du complexe par rapport à celle du peptide est calculé de la manière suivante :

$$\delta\Phi = \overline{\Phi}_W^{cplx} - \overline{\Phi}_W^{pep} \tag{7.3}$$

avec $\overline{\Phi}_W^{cplx}$ et $\overline{\Phi}_W^{pep}$ les potentiels électrostatiques moyens calculés à partir de la simulation du complexe et du peptide respectivement.

Si l'on considère la mutation Ile $\rightarrow$ Arg comme l'insertion d'une charge $+e$, il est alors possible d'estimer la contribution du décalage du potentiel, $\Delta\Delta G_\Phi$, à la différence d'énergie libre de liaison entre les systèmes P^0/P^+ :

$$\Delta\Delta G_\Phi = e(\delta\Phi^0 + \delta\Phi^+)/2 \tag{7.4}$$

avec $\delta\Phi^0$ et $\delta\Phi^+$ la valeur du décalage pour les systèmes neutre et chargé respectivement. La valeur du $\Delta\Delta G_\Phi$ doit ensuite être soustraite à la valeur du $\Delta\Delta G_b$.

7.2 Résultats

7.2.1 Estimation du décalage du potentiel électrostatique

L'utilisation du PME pour traiter les interactions électrostatiques à longue distance entraine un décalage du potentiel électrostatique du système. Ce décalage peut favoriser ou défavoriser les transformations vers un résidu chargé, il est donc nécessaire d'en tenir compte dans les calculs d'énergie libre.

Le potentiel électrostatique dans la région du solvant a été calculé à partir de simulations du peptide Sdc1-F2R et des complexes Tiam1:Sdc1-F2I et Tiam1:Sdc1-F2R. Pour rappel,

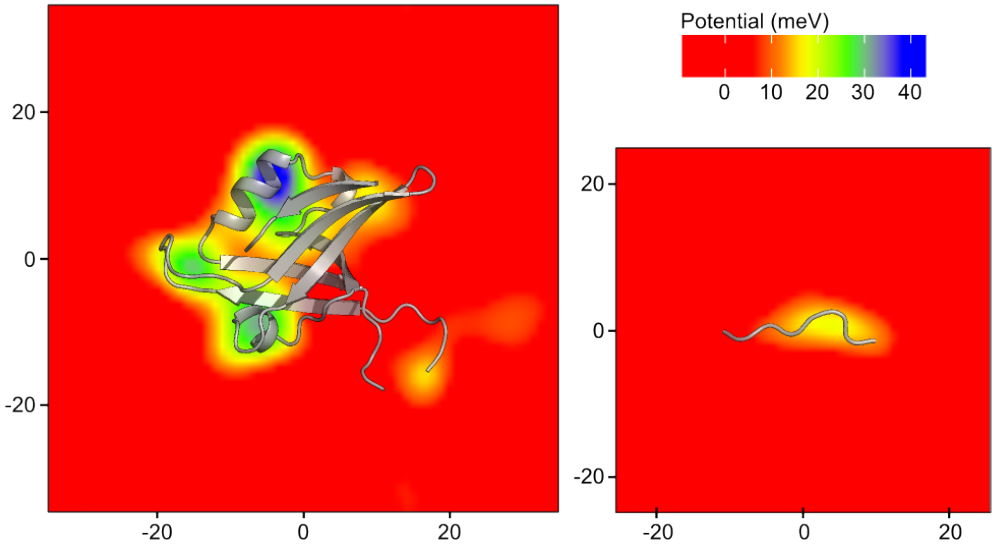


Figure 7.1 – Potentiel électrostatique moyen du complexe Tiam1:P$^+$ et du pepitde isolé. Les potentiels ont été calculés à l'aide de PMEPot. Les valeurs correspondent au potentiel observé au milieu de la boîte. Les dimensions des boîtes sont en Å.

les systèmes chargés et neutres sont respectivement indiqués par P^0 et P$^+$. Pour le variant P$^+$ du peptide, le potentiel du solvant est de 0,3 mEv soit 6×10^{-3} kcal/mol/e. Le volume du peptide étant négligeable par rapport au volume du solvant, sa présence modifie peu le potentiel électrostatique du système qui dépend principalement du solvant. Pour les complexes Tiam1:P^0 et Tiam1:P$^+$, les valeurs obtenues sont respectivement de -19,6 meV et -19,4 meV. Le potentiel électrostatique du solvant étant négatif, cela signifie que celui de la région dans et autour de la protéine est positif et ce indépendamment de la charge du peptide (P^0 ou P$^+$, figure 7.1). En soustrayant la valeur du décalage obtenue pour le peptide à celles obtenues pour les complexes on obtient une valeur moyenne de -19,8 meV ce qui représente une contribution de -0,45 kcal/mol due au déplacement du potentiel électrostatique. Le potentiel électrostatique dans les solutés étant moins négatif lors de la simulation du complexe que du peptide, il est artificiellement plus facile d'effectuer une mutation de charge positive dans le complexe que dans le peptide. La correction est du même ordre de grandeur que les énergies libres de liaison des peptides, sa prise en compte peut donc avoir un impact fort sur la qualité des prédictions. Le décalage du potentiel électrostatique étant moins positif pour les complexes, cette correction doit être soustraite à la valeur du $\Delta\Delta G_b$ lors d'une mutation de charge positive.

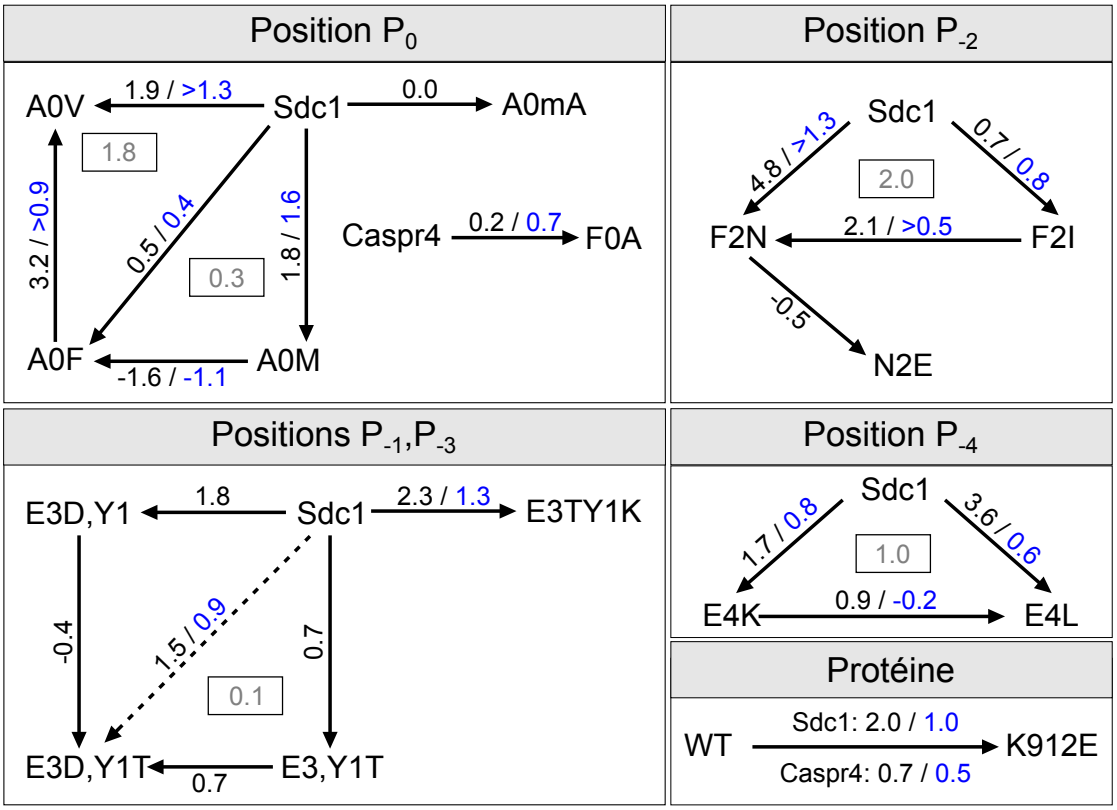


Figure 7.2 – Représentation schématique des transformations alchimiques. Pour chaque position du peptide, les transformations alchimiques effectuées sont représentées par les flèches. À proximité de chaque flèche l'énergie libre de liaisons calculée ainsi que l'énergie libre de liaison expérimentale (lorsqu'elle est disponible) sont indiquées (Calc / Exp). Les erreurs de fermeture associées à chaque cycle thermodynamique correspondent aux valeurs encadrées.

7.2.2 Convergence des calculs d'énergies libres

Les erreurs de fermeture des cycles thermodynamiques sont de bons estimateurs de la convergence des simulations. En effet, si les transformations sont correctement échantillonnées, la somme des énergies libres d'un cycle thermodynamique devrait être nulle quel que soit le champ de force utilisé. Les erreurs de fermeture obtenues pour les six cycles thermodynamiques sont indiquées en figure 7.2 et sont toutes comprises entre 0,0 et 2,0 kcal/mol. Ces résultats sont du même ordre de grandeur que les valeurs obtenues dans des études similaires (Price & Jorgensen [2000] ; Villa *et al.* [2003] ; Dolenc *et al.* [2005]). Ils reflètent la convergence imparfaite de certaines simulations. Une analyse plus poussée permet de mettre en lumière certains défauts d'échantillonnage pouvant expliquer en partie ces résultats.

Dans le cas de P_{-2}, la mutation vers une asparagine provoque un changement conformationnel de l'hélice α_2 suite à l'interaction de N2 avec S903. Bien que ce changement soit de petite ampleur, le passage d'une conformation où l'hélice est dans sa forme native à une forme légèrement courbée n'est probablement pas suffisamment bien échantillonné au cours de la transformation. De la même façon, les dynamiques simples ont montré que la mutation Sdc1-A0V diminuait la stabilité de la partie C-terminale du peptide dans le complexe. La flexibilité de cette région rend l'échantillonnage plus difficile dans les fenêtres où la valine est majoritaire ($\lambda = 0{,}5$ à $\lambda = 1$), ce qui se traduit par des fluctuations plus importantes des valeurs de ΔG intermédiaires entre les différentes séries, de l'ordre de 1 kcal/mol (figure 7.3 A). Cela pourrait expliquer l'erreur de fermeture de cycle de 1,76 kcal/mol obtenue pour le cycle formé par les variants A0, F0 et V0 de Sdc1.

7.2.3 Comparaison des valeurs expérimentales et calculées

Parmi les 22 énergies libres calculées, des valeurs expérimentales d'affinité sont disponibles pour 12 variants. En plus de ces valeurs, la liaison de Sdc1-F2N à Tiam1 est indétectable expérimentalement ce qui indique une énergie libre de liaison au moins 1,3 kcal/mol supérieure à celle de Sdc1 (limite de détection expérimentale). Aucune valine n'étant observée dans la bibliothèque combinatoire à la position P_0 (Shepherd *et al.* [2011]), il est probable que le mutant Sdc1-A0V soit également non liant. Cette comparaison permet à la fois d'estimer les performances du modèle mais également de valider les modèles structuraux. Les valeurs obtenues sont présentées dans le tableau 7.1.

7.2.3.1 Validation des modèles structuraux

La comparaison des énergies libres calculées aux valeurs expérimentales constitue un bon moyen pour valider les modèles structuraux. En effet, si les valeurs sont proches, il y a de fortes chances que le modèle décrive correctement les interactions entre le peptide et la protéine.

Les variants A0F et A0M du peptide Sdc1 sont de parfaits exemples de cette capacité du FEP à discriminer les bons et les mauvais modèles. En effet, dans le cas de ces deux variants, les modèles proposés par SWCRL4 exposent le résidu L915 au solvant ce qui ne correspond ni à l'orientation native, ni à celle observée dans les structures cristallographiques des domaines PDZ dont le résidu à la position P_0 est une Phe ou une Met. Deux modèles alternatifs ont

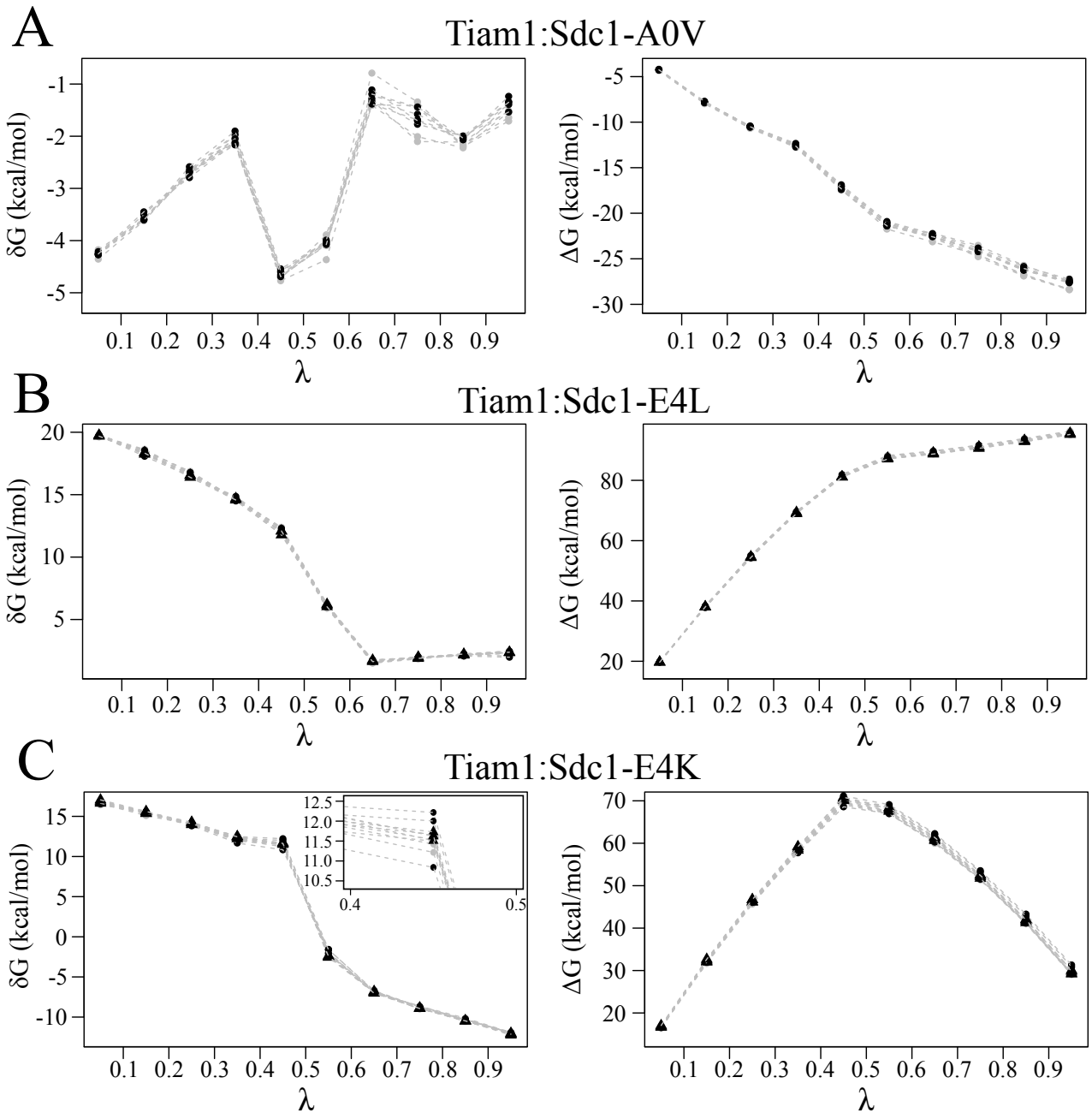


Figure 7.3 – Évolution de l'énergie libre de liaison en fonction de la valeur du paramètre de couplage λ. Les valeurs d'énergie sont indiquées par fenêtre (gauche) ou cumulées (droite). Les valeurs pour les cinq dernières séries sont en noir. Les simulations séquentielles et parallèles sont respectivement représentées par les triangles et les ronds.

Tableau 7.1 – Énergies libres de liaison relatives calculées par transformations alchimiques. Les valeurs indiquées entre paranthèses correspondent aux erreurs statistiques estimées par *bootstrap*.

Mutation	$\Delta\Delta G_b^{Exp}$	$\Delta\Delta G_b^{Calc}$		Err.	$\Delta G^{pep/prot}$		$\Delta G_b^{complexe}$		Corr.[a]
Complexes Sdc1									
A0F	0,43	0,46	(0,96)	0,03	-7,25	(0,05)	-6,79	(0,96)	0,00
A0M	1,56	1,79	(0,45)	0,23	-13,21	(0,04)	-11,42	(0,45)	0,00
A0V	>1,32	1,90	(0,06)	0,00	-27,49	(0,03)	-29,39	(0,05)	0,00
F0M	1,13[b]	1,58	(0,08)	0,45	-2,77	(0,05)	-1,19	(0,06)	0,00
F0V	>0,89[b]	3,20	(0,09)	0,00	-14,80	(0,05)	-11,60	(0,07)	0,00
A0mA	-	0,03	(0,07)	-	126,77	(0,05)	126,80	(0,05)	0,00
F2I	0,80	0,66	(0,09)	-0,14	-0,46	(0,05)	0,20	(0,07)	0,00
F2N	>1,32	4,79	(0,09)	0,00	-6,60	(0,07)	-1,81	(0,06)	0,00
N2E	-	-0,55	(0,13)	-	70,44	(0,12)	70,34	(0,05)	-0,45
I2N	>0.52[b]	2,11	(0,08)	0,00	-9,98	(0,06)	-7,87	(0,05)	0,00
E3D,Y1	-	1,83	(0,25)	-	-1,62	(0,18)	0,21	(0,17)	0,00
E3D,T1	-	0,75	(0,24)	-	-1,46	(0,16)	-0,71	(0,18)	0,00
D3,Y1T	-	-0,37	(0,10)	-	-7,50	(0,07)	-7,87	(0,07)	0,00
E3,Y1T	-	0,74	(0,10)	-	-7,27	(0,07)	-6,53	(0,07)	0,00
E3D,Y1T[c]	0,87	1,48	(0,27)	0,61	-8,93	(0,10)	-7,45	(0,25)	0,00
E3D,Y1T	0,87	-4,54	(0,35)	-5,41	-11,40	(0,11)	-15,94	(0,33)	0,00
E3T,Y1K	1,33	2,34	(0,40)	1,01	36,63	(0,27)	38,07	(0,30)	0,90
E4K	0,81	1,70	(0,38)	0,89	29,06	(0,24)	29,86	(0,30)	0,90
E4L	0,56	3,64	(0,21)	3,08	92,34	(0,14)	95,53	(0,14)	0,45
K4L	-0,25[b]	0,93	(0,15)	1,18	48,66	(0,10)	50,04	(0,11)	-0,45
Complexe Caspr4									
F0A	0,73	0,17	(0,08)	-0,56	8,51	(0,05)	8,68	(0,06)	0,00
Variants de la protéine									
K912E:Sdc1	0,98	2,01	(0,41)	1,03	-44,11	(0,29)	-42,10	(0,28)	0,00
K912E:Caspr4	0,67	0,51	(0,40)	0,16	-44,11	(0,29)	-43,60	(0,27)	0,00

[a] : Correction PME

[b] : Valeurs expérimentales déduites à partir des cycles thermodynamiques (figure 7.2)

[c] : Valeur obtenue par mutations simples en moyennant les deux chemins

donc été produits en contraignant L915 à conserver son orientation cristallographique (figure 7.4). Les énergies de liaison ont été calculées pour les deux modèles et comparées. Les valeurs de $\Delta\Delta G_b$ obtenues pour le peptide Sdc1-A0F sont respectivement de 2,17 kcal/mol et 0,46 kcal/mol lorsque le résidu L915 est exposé et enfoui. La valeur expérimentale étant de 0,43 kcal/mol, le modèle avec L915 enfouie semble être le plus plausible. De la même manière, les valeurs de $\Delta\Delta G_b$ pour le peptide Sdc1-A0M sont de -1,08 et 1,79 kcal/mol lorsque L915 est exposée et enfouie respectivement contre 1,56 kcal/mol expérimentalement. Dans les deux

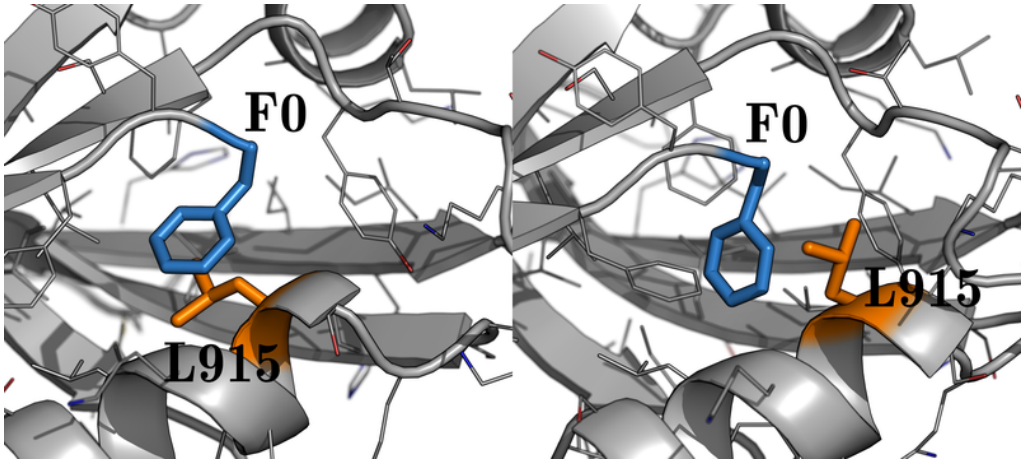


Figure 7.4 – Modèles structuraux du complexe Tiam1:Sdc1$_{A0F}$. Les modèles structuraux avec L915 enfouie et exposée correspondent respectivement aux figures de gauche et de droite.

cas, le FEP montre que les modèles proposés par SCWRL4 possèdent un mauvais rotamère pour L915.

La double mutation Sdc1-E3D,Y1T donne une erreur importante lorsque les transformations sont effectuées simultanément (-5,41 kcal/mol). Les deux positions modifiées étant exposées au solvant, il est peu probable que l'erreur provienne d'une erreur de modélisation. Pour valider cette hypothèse, la double mutation a été effectuée en décomposant la transformation en deux mutations simples. Les valeurs d'énergie obtenues sont alors très proches de la valeur expérimentale (1,48 kcal/mol contre 0,87 kcal/mol expérimentalement), bien que les structures finales soient identiques à celles obtenues lors de la double mutation. Comme nous le verrons plus bas, l'erreur observée dans le cas de la double mutation provient probablement d'un défaut d'échantillonnage des rotamères aux positions P_{-1} et P_{-3}.

Le variant Sdc1-E4L est le seul peptide à présenter une erreur absolue importante par rapport aux valeurs expérimentales (3,08 kcal/mol). Comme nous le verrons plus loin, cette erreur ne remet pas en cause le modèle structural mais serait plutôt due à une erreur liée à l'incapacité du champ de force à décrire correctement la mutation de charge ou à un problème de convergence.

7.2.3.2 Performances du modèle

Parmi les variants étudiés, nous possédons une valeur d'affinité expérimentale pour 12 d'entre eux. Ces données vont permettre de tester les performances prédictives du FEP. L'er-

reur obtenue pour le mutant Sdc1-E4L étant probablement due à un problème de champ de force ou de convergence, ce dernier n'est pas pris en compte par la suite. Excepté pour le peptide Sdc1-E3D,Y1T, seules les transformations directes (sans intermédiaires) sont comparées aux valeurs expérimentales.

Les valeurs moyennes des $\Delta\Delta G_b$ expérimentaux et calculés sont respectivement de 0,82 kcal/mol et 1,24 kcal/mol. Le FEP a donc tendance à sous-estimer l'affinité de 0,42 kcal/-mol en moyenne. L'écart quadratique moyen entre les valeurs expérimentales et calculées est de 0,69 kcal/mol tandis que l'erreur absolue moyenne est de 0,57 kcal/mol. Le coefficient de corrélation de Pearson est de 0,63. Les trois plus grandes erreurs sont de 1,18, 1,03 et 1,01 kcal/mol (Sdc1-K4L, K912E:Sdc1 et Sdc1-E3T,Y1K respectivement) et correspondent à des mutations ioniques. Ces erreurs restent néanmoins dans les valeurs rencontrées avec ce type d'approche (Gilson & Zhou [2007] ; Mikulskis *et al.* [2014]). Les mutations ioniques présentent une erreur moyenne absolue de 0,85 kcal/mol contre 0,34 kcal/mol pour les mutations non ioniques. Afin d'avoir un point de comparaison, nous avons confronté nos résultats à ceux obtenus pour un modèle nul, c'est-à-dire un modèle dans lequel on suppose que tous les peptides sont prédits avec une affinité égale (l'affinité expérimentale moyenne). Ce modèle donne un RMSD de 0,32 kcal/mol, soit une valeur plus faible que celle obtenue par le FEP. Le modèle nul présente cependant un coefficient de corrélation de zéro. Ces résultats sont principalement dus à la faible plage des affinités expérimentales (1,8 kcal/mol) et à la petite taille du jeu de données, le rendant sensible aux erreurs.

Les erreurs statistiques des $\Delta\Delta G$ calculés estimées par *bootstrap* sont comprises entre 0,06 et 0,96 kcal/mol avec une moyenne de 0,24 kcal/mol. Elles sont donc, dans certains cas, supérieures à la plage des affinités expérimentales. Les incertitudes les plus importantes correspondent aux mutations A0F et A0M du peptide Sdc1. L'insertion de ces deux résidus nécessite un réarrangement partiel de l'hélice α_2 ce qui pourrait être responsable de l'incertitude plus importante. La même erreur n'est pas retrouvée pour la mutation Caspr4-F0A, probablement en raison du modèle structural utilisé. En effet, dans ce cas, la forme sauvage possède une phénylalanine à la position P_0 ce qui pourrait faciliter la transformation F→A.

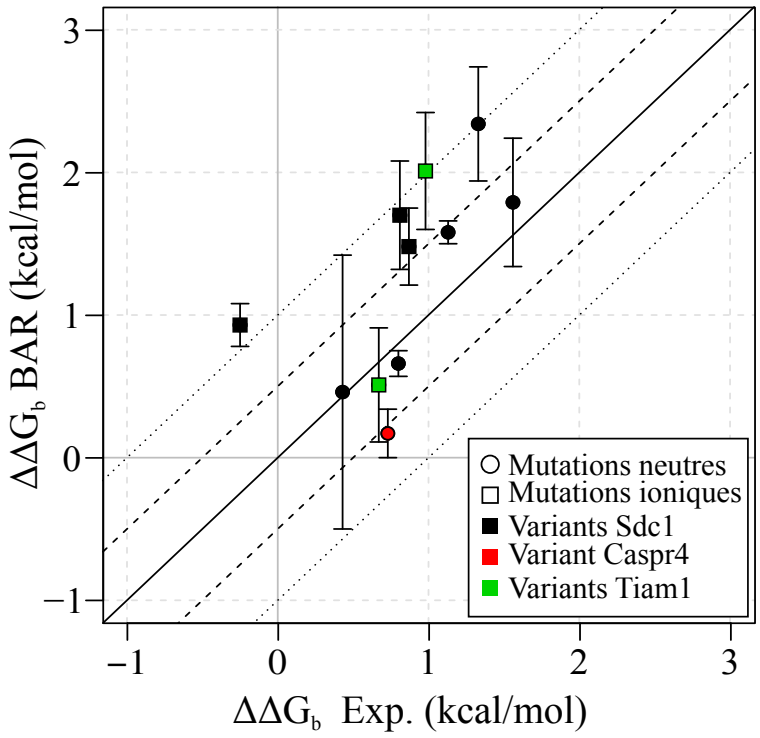


Figure 7.5 – Comparaison des énergies libres de liaison relatives calculées et expérimentales. Les énergies libres de liaison ont été calculées par la méthode BAR. Les barres d'erreurs correspondent à l'erreur calculée par *bootstrap*.

7.2.4 Limites du FEP pour le calcul d'énergie libre de liaison

Parmi les transformations effectuées, la double mutation Sdc1-E3D,Y1T et la mutation ionique Sdc1-E4L sont deux cas pour lesquels l'erreur entre les valeurs calculées et expérimentales sont supérieures à 1 kcal/mol. Pour identifier la cause de ces erreurs des analyses supplémentaires ont été effectuées.

7.2.4.1 Étude de la double mutation Sdc1-E3D,Y1T

Lorsque les deux mutations E3D et Y1T sont effectuées simultanément, l'énergie libre de liaison du variant Sdc1-E3D,Y1T est estimée à -4,54 kcal/mol ce qui est très éloigné de la valeur expérimentale (0,87 kcal/mol). Quand elles sont effectuées successivement, l'erreur est de 0,61 kcal/mol seulement. La transformation pouvant être effectuée de deux manières différentes, les deux possibilités ont été testées. Les résultats sont présentés dans le tableau 7.1. L'ordre des mutations au sein du peptide ne semble pas avoir d'effet sur les énergies

libres de mutation, les valeurs étant de -7,50 et -7,27 kcal/mol pour la position P_{-1} (Y→T) et -1,62 et -1,46 kcal/mol pour la mutation de la position P_{-3} (E→D). Lorsque les mêmes transformations sont effectuées dans le complexe, la mutation Y1T entraine un changement d'énergie libre de -7,87 et -6,53 kcal/mol en présence de D3 et E3 respectivement. De la même façon, la mutation E3D présente une différence d'énergie libre de -0,21 et -0,71 kcal/mol en présence de Y1 et T1 respectivement. Cette différence indique un couplage entre les positions P_{-1} et P_{-3} dans le complexe qui est absent dans le peptide. Il pourrait être expliqué par la présence d'un réseau de liaisons hydrogène entre les chaines latérales des résidus Y1, E3, K6, et N876 (Liu *et al.* [2013]) qui est rompu lorsque D3 est présent en raison de la longueur de sa chaine latérale. Ce couplage se traduit notamment par une énergie de mutation plus favorable à E3 de 1,1 kcal/mol lorsque Y1 est présent et réciproquement. Il est donc probable que le couplage entre P_{-1} et P_{-3} soit mal échantillonné lorsque les deux positions sont modifiées simultanément, ce qui expliquerait l'erreur importante obtenue.

7.2.4.2 Étude des mutants Sdc1-E4K et Sdc1-E4L

La transformation d'un résidu neutre vers un résidu chargé peut modifier la polarisation de la région autour de la mutation. Il est possible que la redistribution des charges ne soit pas correctement décrite par les champs de force additifs. Cela pourrait notamment expliquer les mauvais résultats obtenus pour le mutant Sdc1-E4L. La mutation Sdc1-E4K correspond également à une mutation ionique mais, contrairement à Sdc1-E4L, son énergie libre semble correctement estimée lors des simulations alchimiques. Afin d'identifier les facteurs responsables des énergies libre obtenues pour E4L, des analyses plus poussées ont été effectuées.

Pour s'assurer que les résultats obtenus n'étaient pas le fruit d'un problème de convergence, huit nouvelles séries de 2 ns chacun ont été produites amenant à près de 400 ns le temps de simulation pour la transformation E4L. La prolongation de la simulation, ne change pas la valeur de l'énergie libre (figure 7.3 B). Trois simulations séquentielles ont également été ajoutées à la mutation E4K. Elles mettent en évidence une convergence imparfaite des δG intermédiaires pour les fenêtres autour de $\lambda = 0{,}5$ (figure 7.3 C). L'analyse des structures au cours des trajectoires montre que les mutations E4K et E4L entrainent la perte des interactions E4-R871 et E4-S908 qui sont respectivement présentes 38% et 75% au cours du temps. Dans le cas de la mutation E4K, la transformation entraine également une répulsion des chaines latérales de R871 et K912, ainsi qu'un léger déplacement de l'hélice α_2 permettant à S908

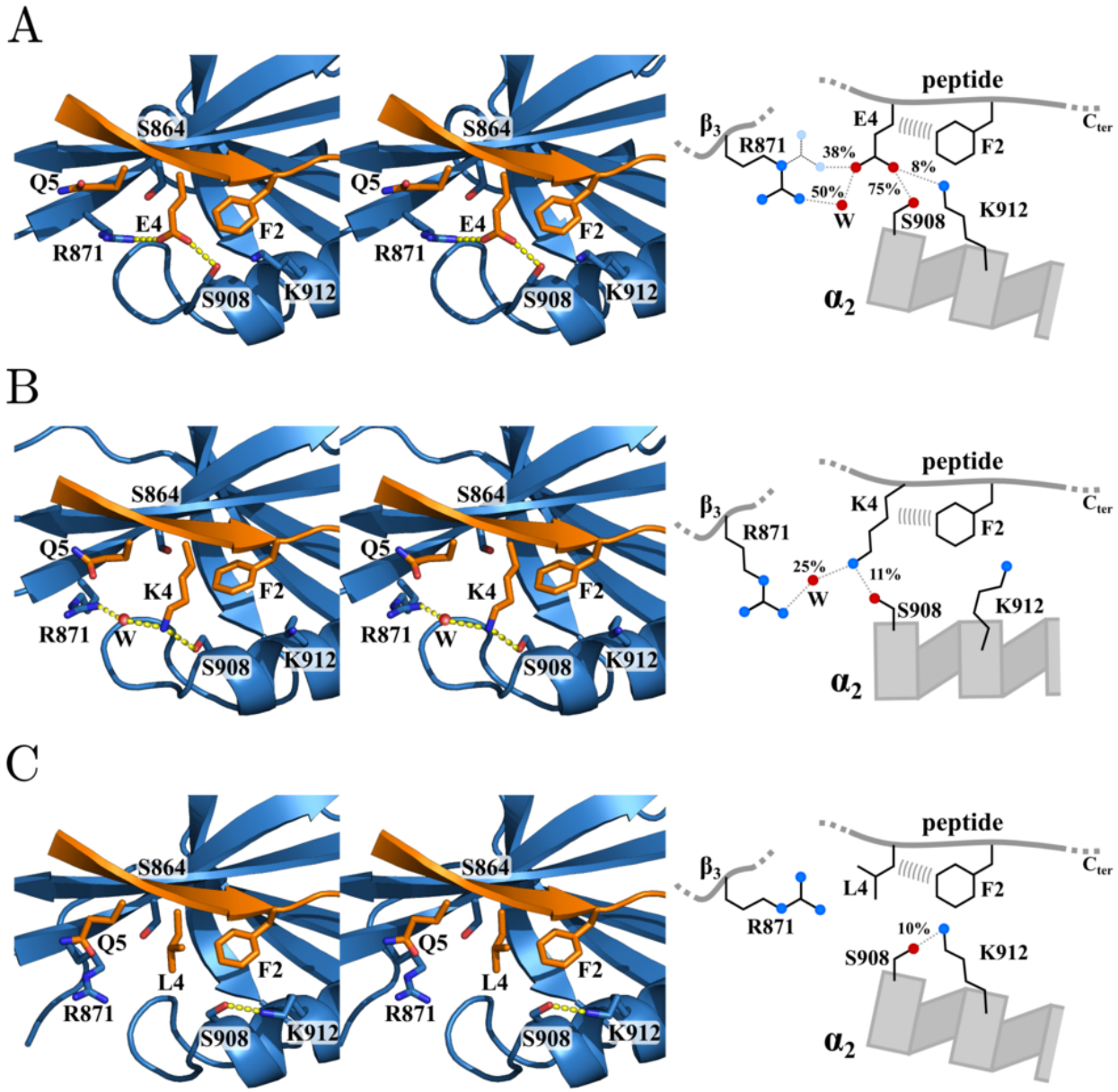


Figure 7.6 – Comparaison des interactions entre les variants de la position P_{-4} du peptide Sdc1. Les complexes Sdc1 natif, E4K et E4L sont respectivement présentés en A, B et C. Pour chaque complexe, les interactions sont présentées en vue stéréo (gauche) et sous forme schématique (droite). La stabilité des interactions (lignes pointillées grises) au cours des simulations est indiquée par les pourcentages. Les conformations alternatives des chaines latérales sont représentés en lignes pointillées noires.

d'interagir ponctuellement avec K4 (figure 7.6). La mutation de la position P_{-4} met donc en jeu des interactions électrostatiques qui pourraient être mal décrites par le champ de force ou mal échantillonnées pendant les transformations.

7.2.5 Prédiction de nouveaux variants

Deux des mutations étudiées ne possèdent pas de valeur expérimentale connue. Le variant Sdc1-F2E est prédit comme non liant puisqu'il possède une valeur de $\Delta\Delta G_b$ supérieure à 1,6 kcal/mol. Le domaine PDZ de Tiam1 appartient à la classe II des domaines PDZ et reconnait donc préférentiellement le motif X-Φ-X-Φ, Φ étant un acide aminé de type hydrophobe. Il est donc probable que le peptide Sdc1-F2E n'est pas reconnu par Tiam1.

La mutation la plus prometteuse est Sdc1-A0mA puisque les résultats FEP indiquent une énergie de liaison égale à celle du peptide sauvage. Cet acide aminé correspond à une alanine pour laquelle le H_α a été remplacé par un groupement méthyle. L'utilisation de cet acide aminé pourrait rendre le peptide résistant aux protéases et donc accroitre sa durée de vie *in vivo*.

7.3 Étude de la forme phosphorylée de Sdc1

7.3.1 Présentation du système

Le peptide Sdc1 possède une tyrosine à la position P_{-1} pouvant être phosphorylée (la forme phosphorylée sera notée pSdc1). Cette phosphorylation joue un rôle important dans la modulation de la signalisation cellulaire et régule, dans le cas de Sdc1, l'adhésion des cellules (Reiland *et al.* [1996] ; Sulka *et al.* [2009]). La résolution de la structure du complexe Tiam1:pSdc1 (4GVC) a mis en évidence un changement dans l'orientation de la tyrosine lorsqu'elle était phosphorylée (Liu *et al.* [2013]). En effet, lorsque la tyrosine n'est pas phosphorylée, elle interagit avec les résidus E3, K6 et N876 à travers un réseau de liaisons hydrogène. Au contraire, lorsqu'elle est phosphorylée, la tyrosine pivote de 90° environ et interagit avec les résidus K879 et T857 au sein d'un sillon formé par l'hélice α_1 et la boucle $\beta_1 - \beta_2$ (figure 7.7). Des analyses par RMN ont montré un couplage fort entre la phosphotyrosine et K879. Malgré ces changements structuraux, la phosphorylation de la tyrosine modifie peu l'affinité de Tiam1 pour Sdc1 ($\Delta\Delta G_b$ = -0,21 kcal/mol).

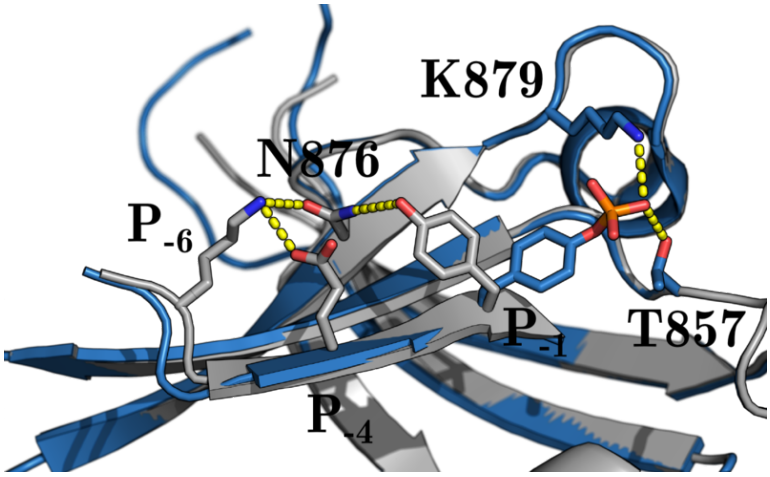


Figure 7.7 – Superposition des structures cristallographiques de Tiam1:Sdc1 et Tiam1:pSdc1. La structure de Tiam1:Sdc1 (4GVD) est représentée en gris tandis que la structure de Tiam1:pSdc1 (4NXR) est en bleu.

7.3.2 Modélisation des complexes

La phosphotyrosine possède un pK_a de 5,96 entre les formes mono et dianioniques de son groupement phosphate (Bienkiewicz & Lumb [1999]). Cette valeur de pK_a étant proche du pH physiologique, il est difficile de déterminer avec certitude l'état de protonation majoritaire dans une protéine. Il est donc nécessaire d'étudier les deux états de protonation afin de prendre en compte toutes les possibilités. Les formes mono et dianioniques de pSdc1 sont modélisées à partir de la structure cristallographique du complexe Tiam1:pSdc1. Des simulations de dynamique moléculaire de 150 ns sont ensuite effectuées en utilisant les mêmes paramètres que précédemment (partie 7.1.2).

7.3.3 Comportement des dynamiques moléculaires

Au cours de la simulation du complexe Tiam1:Sdc1, la tyrosine conserve une orientation proche de celle observée dans la structure cristallographique comme en attestent les valeurs de l'angle χ_1 qui sont majoritairement comprises entre -20 et -90° contre -61° dans le cristal (figure 7.8). Le groupement hydroxyle de Y1 interagit moins de 1% du temps avec la chaine latérale de K879. La forme monoanionique de la phosphotyrosine perd rapidement son interaction avec K879 et présente une orientation proche de l'orientation de la tyrosine non phosphorylée avec un angle χ_1 majoritairement compris entre -45 et -100°. La forme dianionique de la phosphotyrosine conserve une orientation proche de la structure cristallographique pendant 60% de la simulation et interagit avec K879 pendant 82% de la simulation. La phosphotyrosine

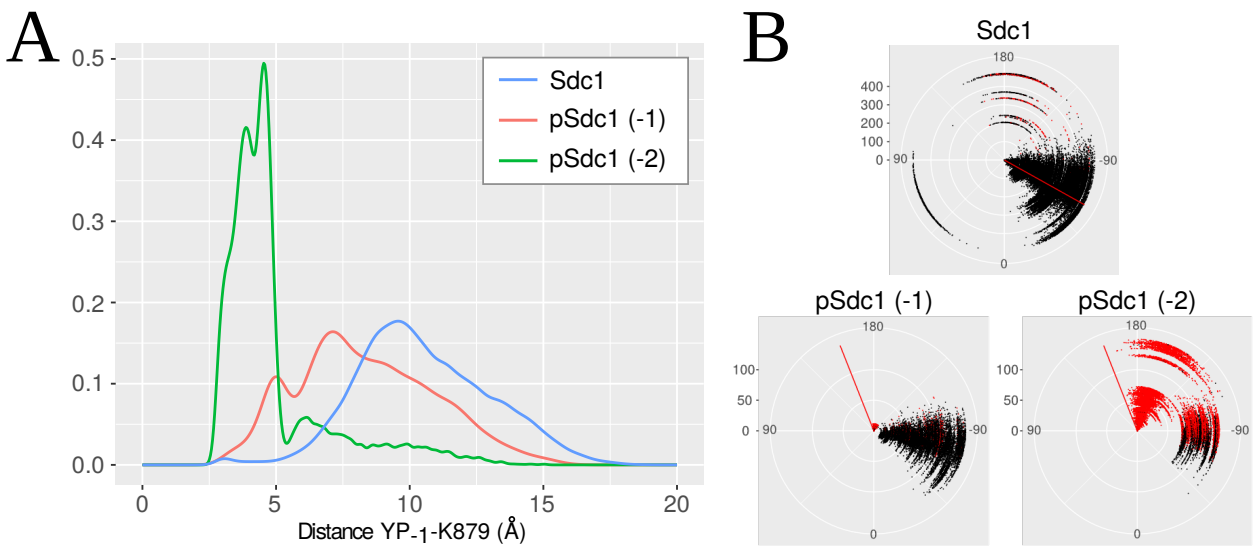


Figure 7.8 – Comportement de la tyrosine P_{-1} au cours des simulations de dynamique moléculaire. A : Distribution de la distance entre le groupement amine de K879 et l'oxygène du groupement hydroxyle (ou de son équivalent dans la forme phosphorylée) au cours des simulations. B : Orientation de l'angle χ_1 de la tyrosine P_{-1}. Les conformations pour lesquelles Y-1 et K879 interagissent sont en rouge.

explore cependant de manière transitoire et réversible une orientation proche de celle de la tyrosine. Contrairement à la forme monoanionique, la phosphotyrosine conserve son interaction avec K879 40% du temps lorsqu'elle s'éloigne de la conformation cristallographique.

Parmi les deux formes de la phosphotyrosine testées, seule la forme dianionique conserve les interactions et l'orientation observées dans le structure cristallographique. Il est donc probable que ce soit la forme majoritaire à pH physiologique puisqu'elle est la seule à même de reproduire les interactions cristallographiques.

7.3.4 Transformations alchimiques

Comme nous avons pu le voir, le calcul d'énergie libre de liaison par transformation alchimique peut être un bon moyen de valider un modèle structural. Pour déterminer si la forme dianionique est effectivement la forme observée *in vitro* cette approche est appliquée aux deux systèmes. L'enjeu de ce test est double puisqu'il permettra d'identifier l'état de protonation de la phosphotyrosine mais également de déterminer s'il est possible d'estimer l'énergie libre de liaison du peptide pSdc1 par transformation alchimique.

L'une des principales difficultés dans le cas présent est que les différentes formes de la tyrosine n'explorent pas les mêmes orientations et n'interagissent pas toutes avec le résidu

K879. Il est donc possible que la transition entre les deux orientations ne soit pas correctement échantillonnée au cours des transformations, menant à des résultats erronés. Afin de pallier le problème, la transformation alchimique est séparée en deux étapes.

La première étape consiste à calculer l'énergie libre de mutation en contraignant les deux chaines latérales du résidu double à adopter l'orientation de la forme dianionique de la phosphotyrosine lors des simulations du complexe. Pour cela, la distance entre le groupement amine de K879 et l'oxygène du groupement hydroxyle (ou de son équivalent dans la forme phosphorylée) est maintenue à une distance maximale de 5 Å par un potentiel semi-harmonique avec une constante de force de 1 kcal/mol/Å^2. Nous avons fait le choix de garder une topologie double y compris pour la transformation entre les deux états de protonation de la phosphotyrosine, en raison des valeurs des charges partielles très différentes entre les atomes des deux formes de la phosphotyrosine.

La contrainte ajoutée, en éloignant les formes phosphorylée monoanionique et non phosphorylée de leur conformation préférentielle, défavorise la liaison de ces deux formes au domaine PDZ par rapport à la forme dianionique. Il est donc nécessaire, dans un second temps, de déterminer l'énergie de contrainte associée aux trois systèmes pour appliquer ensuite un terme correcteur. Pour cela, la contrainte est progressivement relâchée en diminuant la constante de force k (k=$\{1\,;\ 0{,}5\,;\ 0{,}2\,;\ 0{,}075\,;\ 0{,}025\,;\ 0{,}01\,;\ 0\}$ kcal/mol/Å^2). Pour chaque valeur de k, une simulation de 50 ns est effectuée à partir de laquelle les énergies de contraintes sont calculées. L'énergie totale de la contrainte est ensuite estimée à l'aide du BAR. Les résultats obtenus sont présentés dans le tableau 7.2.

Malgré le protocole détaillé utilisé, les valeurs de $\Delta\Delta G_b$ obtenues ne sont pas cohérentes avec les valeurs expérimentales. L'énergie libre de liaison de -12,13 kcal/mol obtenue pour le mutant Y $\rightarrow$ pY$^{(-2)}$ surestime de manière démesurée l'affinité pour la forme phosphorylée. Cette valeur traduit la difficulté qu'a le champ de force à décrire la mutation, probablement en raison des interactions électrostatiques trop importantes causées par l'introduction d'une charge -2. La mutation vers la forme phosphorylée monoanionique présente une énergie libre de liaison de -1,69 kcal/mol soit -1,48 kcal/mol plus faible que la valeur expérimentale. Cette valeur étant plus proche de la valeur expérimentale, cela pourrait signifier que c'est la forme monoanionique qui se lie à Tiam1. Cette hypothèse semble néanmoins contredite par les résultats des simulations simples. L'erreur de fermeture de 3,07 kcal/mol indique que la convergence des simulations n'est pas parfaite, rendant l'interprétation des résultats incertaines.

Tableau 7.2 – Énergies libres calculées pour les transformations alchimiques du peptide pSdc1. Les énergies sont en kcal/mol.

Mutation	Peptide		Complexe		Corr.$_{PME}$	Contraintes		$\Delta\Delta G_b$	
Y → pY$^{(-1)}$	-57,80	(0,17)	-61,69	(0,14)	-0,45	2,65	(0,20)	-1,69	(0,30)
Y → pY$^{(-2)}$	-230,08	(0,73)	-244,81	(0,65)	-0,90	3,50	(0,17)	-12,13	(1,00)
pY$^{(-2)}$ → pY$^{(-1)}$	174,00	(0,58)	181,76	(0,64)	0,45	-0,84	(0,11)	7,37	(0,87)

Les mutations de phosphorylation sont connues pour être difficilement traitables par les méthodes de perturbation d'énergie libre, en partie à cause de leur énergie de solvatation importante (Gumbart *et al.* [2013]). D'autres méthodes comme le potentiel de force moyenne (PMF pour *Potential of Mean Force*) ont déjà été appliquées avec succès pour le calcul de l'énergie libre de liaison absolue de peptides phosphorylés (Woo & Roux [2005] ; Buch *et al.* [2011]). Cette méthode, qui simule explicitement le processus de liaison du peptide, pourrait être appliquée.

7.3.5 Estimation du décalage du pK_a de la phosphotyrosine dans la protéine

À partir des transformations alchimiques des formes mono et dianionique de la phosphotyrosine, la différence de pK_a du phosphate dans la protéine et en solution peut être estimée. La méthode repose sur le cycle thermodynamique présenté en figure 7.9. La constante d'équilibre K_a de la liaison du proton dans les deux états est reliée à l'énergie libre standard ΔG par (Simonson *et al.* [2004]) :

$$\Delta G = -kT \log K_a \tag{7.5}$$

on a alors

$$\mathrm{p}K_a = -\log_{10} K_a = \frac{1}{2.303kT}\Delta G \tag{7.6}$$

et

$$\mathrm{p}K_{a,prot} = \mathrm{p}K_{a,ref} + \frac{1}{2.303kT}\Delta\Delta G \tag{7.7}$$

où p$K_{a,prot}$ correspond au pK_a de la phosphotyrosine dans le complexe, p$K_{a,ref}$ au pK_a de la phosphotyrosine en solution (p$K_{a,ref} = 5{,}96$, Bienkiewicz & Lumb [1999]), et $\Delta\Delta G = \Delta G_{prot} - \Delta G_{pep}$ est la double différence d'énergie libre du cycle thermodynamique en figure 7.9.

Le $\Delta\Delta G$ associé à la déprotonation de la phosphotyrosine est de 7,37 kcal/mol. Le pK_a obtenue pour la phosphotyrosine dans le complexe est donc de 0,56 ce qui signifie qu'à pH

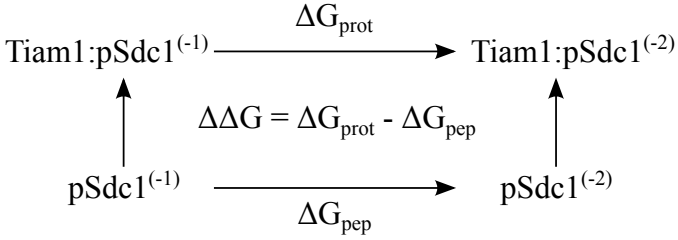


Figure 7.9 – Cycle thermodynamique pour le calcul du décalage du pK_a. La flèche supérieure représente la liaison d'un proton sur le groupement phosphate de la phosphotyrosine dans le complexe. La flèche inférieure représente la liaison d'un proton sur le groupement phosphate de la phosphotyrosine dans le peptide.

physiologique seule la forme dianionique est présente dans le complexe. Cet état de protonation serait stabilisé par les interactions avec T857 et R879. Ce résultat est en accord avec les données de simulations qui montrent que seule cette forme est capable de conserver une orientation proche de la structure cristallographique.

7.4 Discussion et conclusions

Le but principal de cette étude était de tester les performances de la méthode de transformation alchimique dans le cadre du calcul de l'énergie libre de liaison relative. Cette approche a été largement utilisée ces dernières années et a connu de nombreux succès. Les résultats obtenus pour les 12 peptides possédant des données expérimentales sont en accord avec ces dernières (excepté pour Sdc1-E4L) et permettent notamment d'obtenir un RMSD de 0,69 kcal/mol et un coefficient de corrélation de 0,63. Ces performances sont très proches de celles obtenues avec le modèles PB/LIE optimisé sur 35 peptides (chapitre 6). Le FEP semble cependant capable de discriminer les peptides non-liants ($\Delta\Delta G_b \geq 1{,}6$ kcal/mol) ce qui n'était pas le cas du modèle PB/LIE. Trois ingrédients principaux sont indispensables à la bonne réussite de cette approche : la qualité des modèles initiaux, celle de l'échantillonnage, et la capacité du champ de force à décrire les interactions.

Les domaines PDZ présentent l'avantage d'avoir un mode de liaison bien connu. Néanmoins, les variants Sdc1-A0F et Sdc1-A0M sont deux exemples permettant de démontrer que le mauvais positionnement d'une seule chaine latérale peut suffire à obtenir des résultats erronés. Ces exemples montrent également l'impact que peut avoir le choix de la structure initiale dans l'étape de modélisation sur la qualité du modèle mais également sur la qualité des affinités calculées. En effet, bien que donnant une valeur très proche de la valeur expérimentale,

le variant Sdc1-A0F est entaché d'une incertitude importante, du même ordre de grandeur que les affinités expérimentales, ce qui n'est pas le cas du variant Caspr4-F0A qui correspond pourtant à la même transformation. La seule différence entre ces deux systèmes provient du modèle structural utilisé : Tiam1:Sdc1 pour Sdc1-A0F et QM:Caspr4 pour Caspr4-F0A. Le peptide Caspr4 possède déjà une Phe à la position P_0, ce qui n'est pas le cas de Sdc1 qui possède un résidu Ala. L'introduction de la Phe dans le cas du variant Sdc1-A0F entraine donc des changements structuraux plus importants au niveau de l'hélice α_2 qui sont responsables de l'incertitude plus élevée.

La qualité de l'échantillonnage constitue le deuxième point important dans l'obtention de valeurs d'énergies libres justes et ceci est d'autant plus vrai lorsque la transformation entraine des changements conformationnels. Une des méthodes permettant d'étudier la convergence des transformations consiste à calculer les erreurs de fermeture des cycles thermodynamiques. Cette approche n'est pas toujours possible puisqu'elle nécessite que les différents ligands soient relativement similaires. Dans notre cas, les erreurs de fermeture supérieures à 1 kcal/mol laissent toutefois penser que ces résultats pourraient être encore améliorés, notamment par un meilleur échantillonnage des changements conformationnels, les principales erreurs rencontrées provenant de la flexibilité importante du système. Une amélioration possible serait de réduire l'entropie conformationnelle en imposant un certain nombre de contraintes, à la manière de ce qui a été fait pour la phosphotyrosine, puis dans un second temps de déterminer la contribution de ces contraintes à l'énergie libre de liaison (Gumbart *et al.* [2013]). Cette approche pourrait être utilisée dans le cas de la mutation Sdc1-A0V pout limiter les fluctuations de l'extrémité C-terminale du peptide. D'autres méthodes comme la dynamique accélérée avec échange de répliques (FEP-REMD ou encore H-REMD pour *Hamiltonian Replica-exchange Molecular Dynamics*) ont également été développées dans l'optique d'échantillonner plus efficacement l'espace conformationnel (Jiang *et al.* [2009] ; Jiang & Roux [2010]). Le cas de la double mutation E3D,Y1T est intéressant puisqu'il permet de montrer que le défaut d'échantillonnage, y compris au niveau des rotamères, peux entrainer des erreurs de plusieurs kcal/mol. Cela avait déjà été observé dans des études précédentes (Mobley *et al.* [2007a,b]). Dans le cas où la double mutation fait intervenir des résidus couplés, il semble donc préférable de décomposer la transformation en deux mutations simples.

Parmi les 12 mutations pour lesquelles des données sont disponibles, la moitié correspond à des mutations ioniques (apparition ou inversion de charge). Toutes ces mutations sont cor-

rectement prédites, avec une erreur maximale de 1 kcal/mol, excepté le mutant Sdc1-E4L. Les champs de force additifs sont connus pour surestimer les interactions électrostatiques entre les groupes polaires. Cela pourrait expliquer les bons résultats obtenus pour les mutations d'inversion de charge. En effet, les résidus des états initiaux et finaux étant tous deux polaires, les deux états sont favorisés, ce qui permet de compenser l'erreur introduite par le champ de force. Cette compensation n'est pas possible dans le cas de Sdc1-E4L ce qui pourrait expliquer l'erreur importante observée. Cependant, les bons résultats obtenus pour le mutant Sdc1-K4L ne confortent pas cette hypothèse. Les différences d'énergies libres expérimentales étant très petites (0,56 kcal/mol dans le cas de Sdc1-E4L) et les énergie libre de mutation parfois très grandes (95 kcal/mol pour le même mutant lié à Tiam1), il est également possible que la méthode soit limitée par la précision statistique du FEP. Un problème identique avait déjà été relevé dans l'étude de la liaison du peptide phosphorylé au domaine SH2 de la protéine p56 (Woo & Roux [2005]), la transformation géométrique donnant dans ce cas de meilleurs résultats. Ce cas de figure, très similaire à celui de pSdc1, laisse penser qu'une telle approche pourrait être utilisée dans notre cas. Pour ces variants, il serait intéressant de tester d'autres champs de force et notamment un champ de force polarisable tel que DRUDE (Lamoureux *et al.* [2003] ; Lamoureux & Roux [2003]) capable de modéliser la redistribution des charges et donc plus à même de décrire les mutations ioniques. L'utilisation d'un autre champ de force additif pourrait également assurer la reproductibilité des résultats (Simonson *et al.* [2002]). Enfin, il est intéressant de noter que la plupart des transformations surestiment l'énergie libre de liaison en faveur du type natif. Cela pourrait être dû au modèle structural utilisé (le complexe WT:Sdc1) qui n'aurait pas suffisamment le temps de se relaxer au cours des transformations pour accueillir le type muté. Des simulations plus longues pourraient améliorer la convergence des résultats.

Parmi les variants Sdc1 pour lesquels nous n'avions pas de données expérimentales, seule la diméthyle alanine en position P_0 ne détériorer pas l'affinité. Si ce résultat s'avère être exact, l'utilisation d'un acide aminé non naturel pourrait être intéressante dans la confection de peptides inhibiteurs car une telle modification pourrait les protéger des protéases.

Les transformations alchimiques restant couteuses en termes de temps de calculs, leur utilisation est difficilement applicable à la recherche haut débit de nouveaux peptides inhibiteurs. Mais couplée avec une approche semi-empirique comme le MM/PBSA, elle peut constituer une étape de raffinement supplémentaire avant les tests expérimentaux.

Conclusion

Les interactions protéine-protéine (PPI) jouent un rôle majeur au sein des cellules et sont, de ce fait, des cibles thérapeutiques potentielles. La reconnaissance entre partenaires est généralement assurée par des petits domaines protéiques spécialisés. Au cours de cette thèse nous nous sommes intéressés au domaine PDZ de la protéine Tiam1 qui est impliquée dans des processus cellulaires aussi variés que la migration, l'adhésion ou encore la croissance cellulaire. Des formes mutées de Tiam1 étant retrouvées dans certains cancers, cette protéine pourrait constituer une bonne cible thérapeutique. L'identification et la conception d'inhibiteurs des PPI par des approches expérimentales reste toutefois un processus long et couteux. L'outil informatique se trouve donc être une excellente alternative puisqu'il permet d'étudier et de caractériser des complexes protéiques rapidement et à grande échelle.

Le but de cette thèse était de caractériser, par des approches de dessin computationnel de protéine, de simulation de dynamique moléculaire et de calcul d'énergie libre, des complexes Tiam1:peptide. Cette étude visait à comprendre les mécanismes impliqués dans la spécificité de Tiam1 pour ses partenaires et, le cas échéant, à proposer des peptides inhibiteurs.

Dans un premier temps, nous avons utilisé le programme de CPD Proteus pour explorer les séquences compatibles avec le pli des domaines PDZ. Le modèle déplié a tout d'abord été paramétré pour reproduire les fréquences en acides aminés des séquences naturelles de domaines PDZ (chapitre 3). Deux nouveautés ont été introduites dans la procédure d'optimisation : la séparation des positions en deux partitions (exposé/enfoui) et l'optimisation des énergies de référence par groupes et non plus par types. Une fois le jeu de paramètres optimisé, la qualité des séquences Proteus produites à partir des squelettes de Tiam1 et Cask est comparable à celle des séquences Rosetta. Ces séquences sont également proches des séquences naturelles des domaines PDZ. L'exploration de quatre positions impliquées dans la spécificité de Tiam1 a montré qu'il était possible d'extraire des informations qualitatives sur l'affinité des complexes PDZ:peptide à partir des résultats Proteus.

Afin d'étudier la stabilité des séquences Proteus, nous avons effectué des simulations de dynamique moléculaire pour dix d'entre elles (chapitre 4). Toutes les séquences conservent leurs structures secondaires au cours des simulations et deux des trois séquences prolongées à 1 μs restent stables. La comparaison des simulations avec des données RMN a mis en évidence une flexibilité accrue des boucles et de l'hélice α_2 dans les simulations. Des analyses *in vitro* ont ensuite montré qu'une des séquences Proteus ainsi que trois de ses variants étaient partiellement repliés 50°C. La stabilité de ces séquences pourrait probablement être encore améliorée en modifiant quelques positions. La paramétrisation des énergies de référence et l'étude des séquences générées par Proteus a fait l'objet d'une publication en co-premier auteur (Mignon *et al.* [2017]).

Dans la seconde partie, nous nous sommes intéressés aux méthodes de calcul d'énergies libres pour prédire l'affinité des complexes Tiam1:peptide. Les énergies libres de liaison ont tout d'abord été estimées en utilisant une classe de modèles semi-empiriques (chapitre 6). Cette approche présente l'avantage de pouvoir être appliquée à un grand nombre de complexes mais nécessite une première phase de paramétrisation. Deux difficultés sont la flexibilité importante du système et la faible plage des valeurs d'affinité. Les modèles ont été paramétrés en utilisant 37 complexes pour lesquels des données expérimentales d'affinité étaient disponibles. Le modèle PB/LIE permet d'obtenir une erreur absolue moyenne de 0,4 kcal/mol et un coefficient de corrélation de 0,64 entre les valeurs expérimentales et calculées. L'approche mono-trajectoire utilisée présente cependant quelques limites puisqu'elle traite implicitement les changements conformationnels suite à la liaison du peptide. Cela nous a contraint à exclure certains complexes du modèle et à introduire un terme correcteur aux variants du QM. Cette approche reste cependant moins lourde à mettre en place que les approches à deux ou trois trajectoires qui, dans notre cas, n'ont pas donné de meilleurs résultats. Une publication portant sur cette étude a récemment été acceptée (Panel *et al.* [2017]).

Nous avons ensuite évalué l'énergie libre de liaison d'un plus petit nombre de complexes par une méthode plus rigoureuse, le FEP, qui n'utilise aucun paramètre ajustable (chapitre 7). Parmi les 22 transformations alchimiques effectuées, 12 possèdent des valeurs expérimentales. Le FEP prédit l'énergie libre de liaison avec une erreur absolue moyenne de 0,6 kcal/mol, en excluant un mutation. Seule la mutation Sdc1-E4L présente une erreur supérieure à 1 kcal/mol. Ce résultat pourrait probablement être amélioré en allongeant le temps de simulation ou en traitant plus rigoureusement les interactions électrostatiques. En effet, le champ de force

polarisable DRUDE donne une erreur nettement réduite de 0,9 kcal/mol seulement pour Sdc1-E4L. Ces résultats feront l'objet d'une publication en cours de préparation.

Malgré les nombreux variants peptidiques testés au cours de cette étude, aucun inhibiteur de Tiam1 n'a pu être identifié. Cependant, l'introduction d'un acide aminé non naturel à l'extrémité C-terminale du peptide pourrait augmenter sa stabilité *in vivo* sans modifier son affinité pour Tiam1. De plus, les résultats FEP ont montré qu'il peut exister un couplage entre les positions du peptide. Une combinaison de mutations ponctuelles pourrait alors le rendre plus affin. L'affinité n'est cependant pas la seule donnée à prendre en compte. En effet, il est également nécessaire que l'inhibiteur soit spécifique de la cible et biodisponible.

Annexe A

Optimisation des énergies de référence

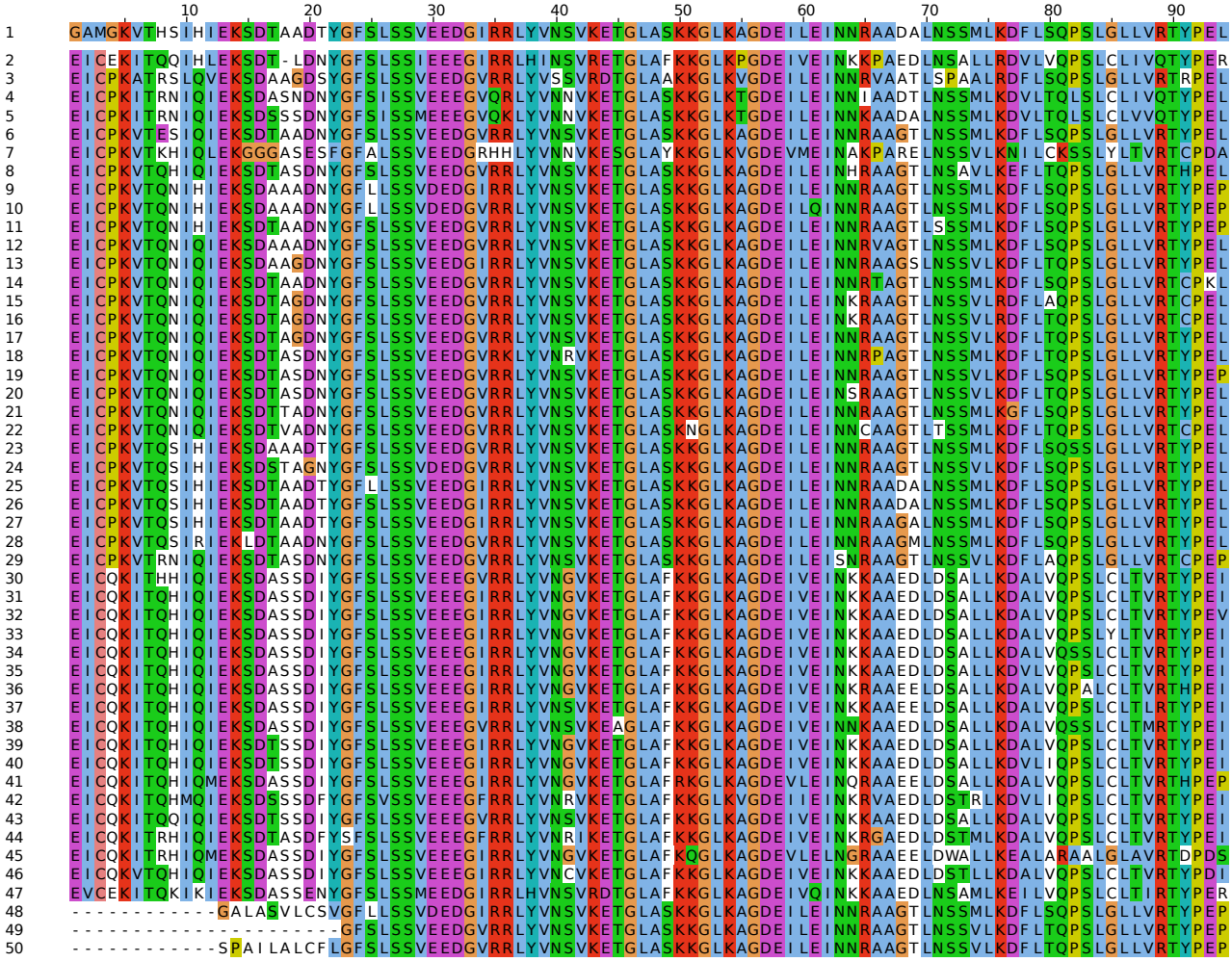


Figure A.1 – Séquences homologues à Tiam1 utilisées lors de l'optimisation des énergies de référence. Le pourcentage d'identité de séquence avec Tiam1 est compris entre 60 et 85%.

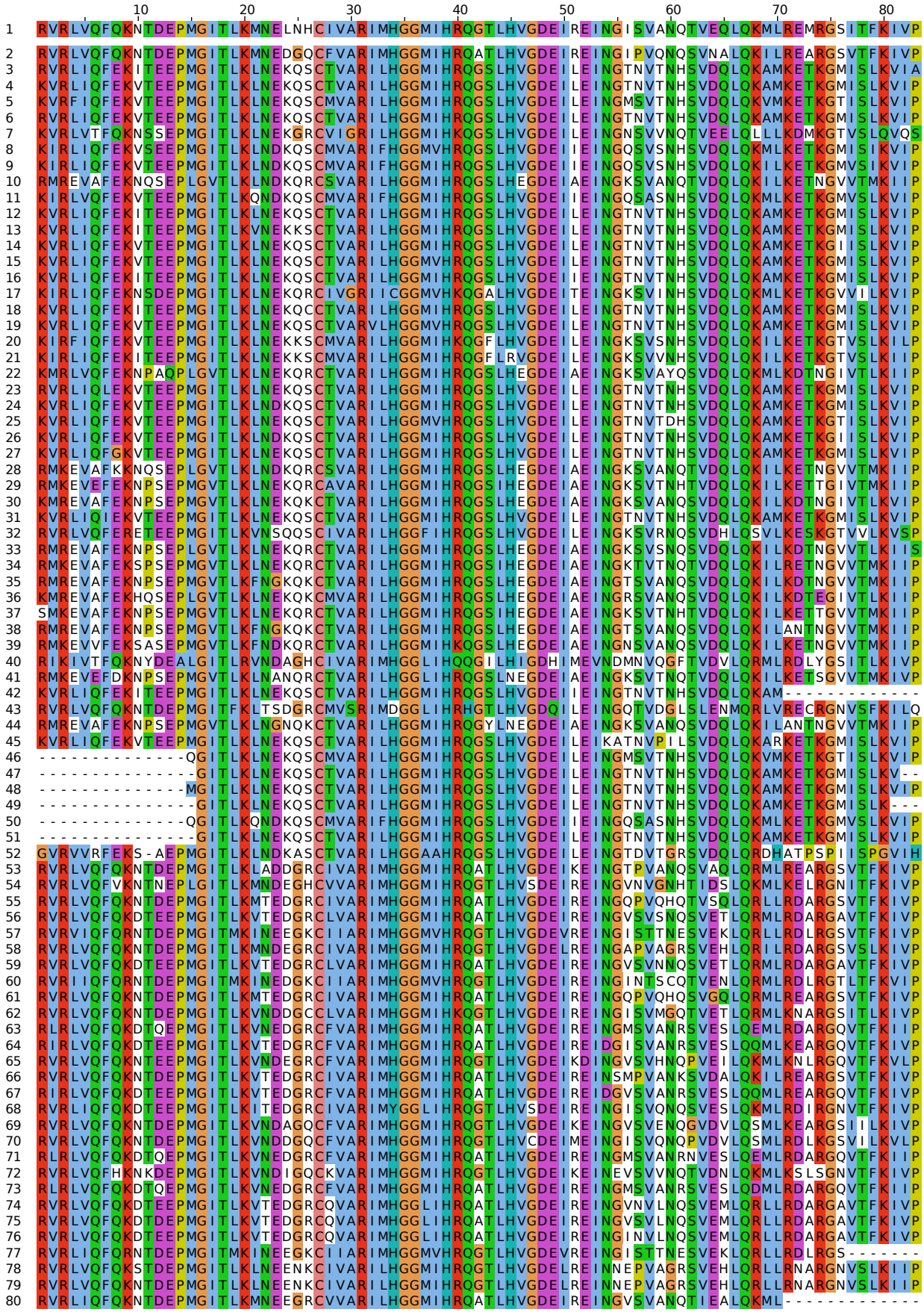


Figure A.2 – Séquences homologues à Cask utilisées lors de l'optimisation des énergies de référence. Le pourcentage d'identité de séquence avec Cask est compris entre 60 et 85%.

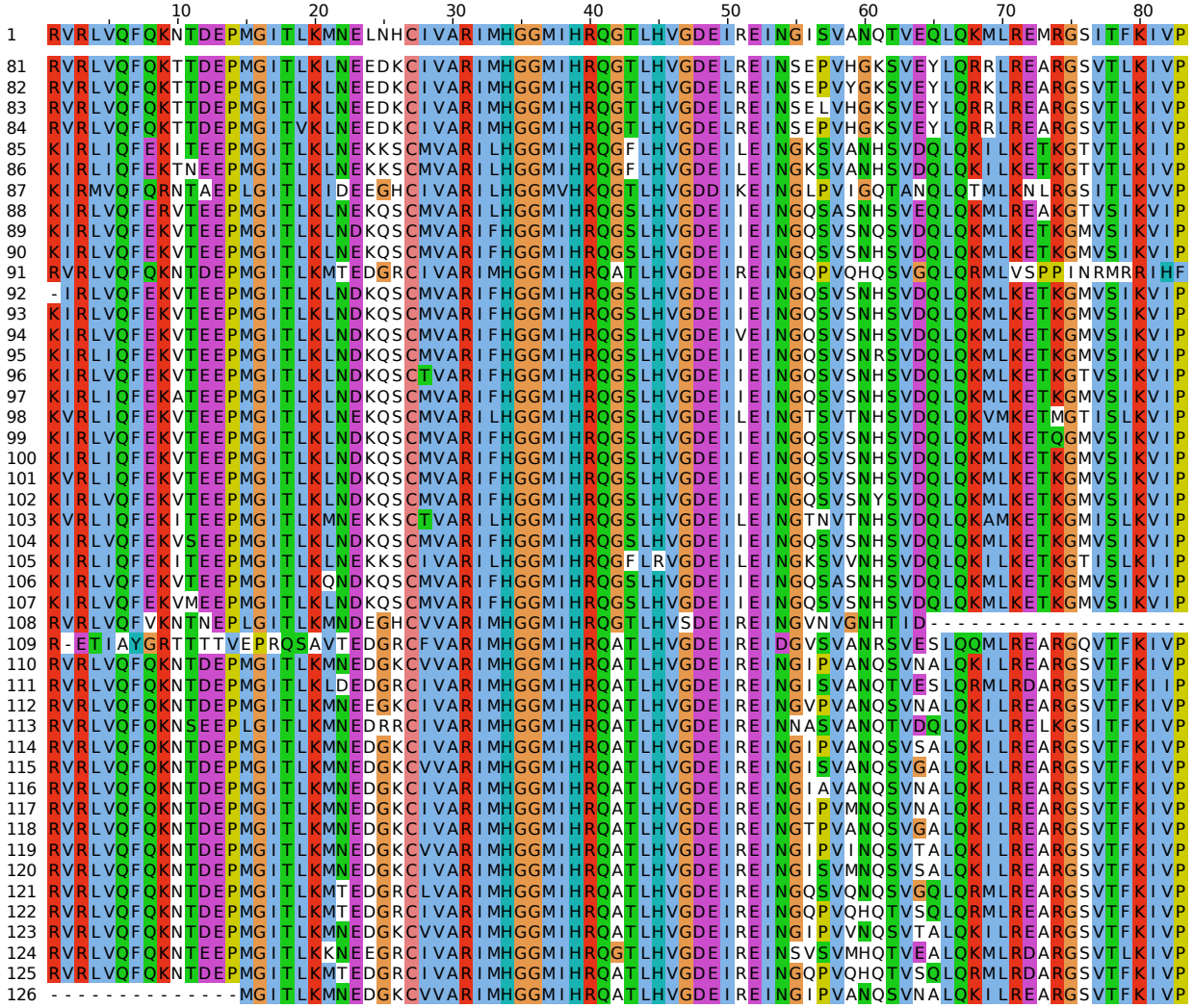


Figure A.2 (suite) - Séquences homologues à Cask utilisées lors de l'optimisation des énergies de référence. Le pourcentage d'identité de séquence avec Cask est compris entre 60 et 85%.

Annexe B

Sélection des séquences Proteus pour le test de stabilité

Figure B.1 – Prédiction des structures secondaires pour les séquences générées par Proteus. Les structures secondaires ont été prédites à l'aides des programmes Porter4.0 et YASPIN. Les structures secondaires présentent dans la structure de Tiam1 ont été déterminée à partir de la structure cristallographique 4GVD à l'aide du programme DSSP. Les hélices, les brins et les coudes sont respectivement représentés par les lettre "H","E" et "T"

Bibliographie

Aarts M. (2002). Treatment of ischemic brain damage by perturbing NMDA receptor- PSD-95 protein interactions. *Science* **298**, 846–850.
cité page 8

Adam L., Vadlamudi R., McCrea P. & Kumar R. (2001). Tiam1 overexpression potentiates heregulin-induced lymphoid enhancer factor-1/β-catenin nuclear signaling in breast cancer cells by modulating the intercellular stability. *Journal of Biological Chemistry* **276**, 28443–28450.
cité page 14

Aksimentiev A. & Schulten K. (2005). Imaging alpha-hemolysin with molecular dynamics: ionic conductance, osmotic permeability, and the electrostatic potential map. *Biophys J* **88**, 3745–3761.
cité page 151

Aldeghi M., Heifetz A., Bodkin M., Knapp S. & Biggin P. (2016). Accurate calculation of the absolute free energy of binding for drug molecules. *Chemical Science* **7**, 207–218.
cité page 146

Alford R.F., Leaver-Fay A., Jeliazkov J.R., O'Meara M.J., DiMaio F.P., Park H., Shapovalov M.V., Renfrew P.D., Mulligan V.K., Kappel K., Labonte J.W., Pacella M.S., Bonneau R., Bradley P., Dunbrack R.L., Das R., Baker D., Kuhlman B., Kortemme T. & Gray J.J. (2017). The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation* **13**, 3031–3048.
cité page 36

Allouche D., André I., Barbe S., Davies J., Givry S., Katsirelos G., O'Sullivan B., Prestwich S., Schiex T. & Traoré S. (2014). Computational protein design as an optimization problem. *Artificial Intelligence* **212**, 59–79.
cité page 34

Almlöf M., Carlsson J. & Åqvist J. (2007). Improving the accuracy of the linear interaction energy method for solvation free energies. *Journal of Chemical Theory and Computation* **3**, 2162–2175.
cité page 109

Altschul S., Gish W., Miller W., Myers E. & Lipman D. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.
cité page 42

Amacher J., Zhao R., Spaller M. & Madden D. (2014). Chemically modified peptide scaffolds target the CFTR-associated ligand PDZ domain. *PLOS ONE* **9**.
cité page 8

Andrew Leaver-Fay A., O'Meara M., Tyka M., Jacak R., Song Y., Kellogg E., Thompson J., I.W. D., R.A. P., Lyskov S., Gray J., Kortemme T., Richardson J., Havranek J., Snoeyink J., Baker D. & Kuhlman B. (2013). Scientific benchmarks for guiding macromolecular energy function improvement. In *Methods in Enzymology*, 109–143 (Elsevier).
cité page 36

Appleton B., Zhang Y., Wu P., Yin J., Hunziker W., Skelton N., Sidhu S. & Wiesmann C. (2006). Comparative structural analysis of the erbin PDZ domain and the first PDZ domain of ZO-1. *Journal of Biological Chemistry* **281**, 22312–22320.
cité page 6

Åqvist J. & Hansson T. (1996). On the validity of electrostatic linear response in polar solvents. *Journal of Physical Chemistry* **100**, 9512–9521.
cité page 109

Åqvist J. & Marelius J. (2001). The linear interaction energy method for predicting ligand binding free energies. *Combinatorial Chemistry & High Throughput Screening* **4**, 613–626.
cité page 110

Åqvist J., Medina C. & Samuelsson J. (1994). A new method for predicting binding affinity in computer-aided drug design. *Protein Engineering, Design and Selection* **7**, 385–391.
cité page 108

Åqvist J., Luzhkov V. & Brandsdal B. (2002). Ligand binding affinities from md simulations. *Accounts of Chemical Research* **35**, 358–365.
cité page 108

Bach A., Eildal J., Stuhr-Hansen N., Deeskamp R., Gottschalk M., Pedersen S., Kristensen A. & Stromgaard K. (2011). Cell-permeable and plasma-stable peptidomimetic inhibitors of the postsynaptic density-95/n-methyl-d-aspartate receptor interaction. *Journal of Medicinal Chemistry* **54**, 1333–1346.
cité page 8

Bach A., Clausen B., Moller M., Vestergaard B., Chi C., Round A., Sorensen P., Nissen K., Kastrup J., Gajhede M., Jemth P., Kristensen A., Lundstrom P., Lambertsen K. & Stromgaard K. (2012). A high-affinity, dimeric inhibitor of PSD-95 bivalently interacts with PDZ1-2 and protects against ischemic brain damage. *Proceedings of the National Academy of Sciences* **109**, 3317–3322.
cité page 8

Baldwin E., Hajiseyedjavadi O., Baase W. & Matthews B. (1993). The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science* **262**, 1715–1718.
cité page 25

Baruch H. & Wendell L. (2001). Mechanism and role of PDZ domains in signaling complex assembly. *Journal of Cell Science* **114**, 3219–3231.
cité page 4

Basdevant N., Weinstein H. & Ceruso M. (2006). Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. *Journal of the American Chemical Society* **128**, 12766–12777.
cité page 10

Ben-Shalom I., Pfeiffer-Marek S., Baringhaus K. & Gohlke H. (2017). Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations. *Journal of Chemical Information and Modeling* **57**, 170–189.
cité page 114

Bennett C. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **22**, 245–268.
cité pages 105 et 148

Best R. & Hummer G. (2009). Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *Journal of Physical Chemistry B* **113**, 9004–9015.
cité page 83

Beutler T., Mark A., van Schaik R., Gerber P. & van Gunsteren W. (1994). Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters* **222**, 529–539.
cité page 148

Beveridge D. & DiCapua F. (1989). Free energy via molecular simulation: Applications to chemical and biomolecular systems. *Annual Review of Biophysics and Biophysical Chemistry* **18**, 431–492.
cité pages 101 et 145

Bhattacherjee A. & Wallin A. (2013). Exploring protein-peptide binding specificity through computational peptide screening. *PLOS Computational Biology* **9**, e1003277.
cité page 11

Bienkiewicz E. & Lumb K. (1999). Random-coil chemical shifts of phosphorylated amino acids. *Journal of Biomolecular NMR* **15**, 203–206.
cité pages 163 et 166

Blöchliger N., Xu M. & Caflisch A. (2015). Peptide binding to a PDZ domain by electrostatic steering via nonnative salt bridges. *Biophysical Journal* **108**, 2362–2370.
cité page 9

Boissier P. & Huynh-Do U. (2014). The guanine nucleotide exchange factor Tiam1: A janus-faced molecule in cellular signaling. *Cellular Signalling* **26**, 483–491.
cité pages 13 et 14

Born M. (1920). Volumen und hydratationswarme der ionen. *Z. Phys.* **1**, 45–48.
cité page 30

Bowman G. (2016). Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *Journal of Computational Chemistry* **37**, 558–566.
cité page 88

Brandsdal B., Osterberg F., Almlöf M., Feierberg I., Luzhkov V. & Åqvist J. (2003). Free energy calculations and ligand binding. *Advances in protein chemistry* **66**, 123–158.
cité page 125

Brooks B.R., Brooks III C.L., Mackerell A.D., Nilsson L., Petrella R.J., Roux B., Won Y., Archontis G., Bartels C., Boresch S., Caflisch A., Caves L., Cui Q., Dinner A.R., Feig M., Fischer S., Gao J., Hodoscek M., Im W., Kuczera K., Lazaridis T., Ma J., Ovchinnikov V., Paci E., Pastor R.W., Post C.B., Pu J.Z., Schaefer M., Tidor B., Venable R.M., Woodcock H.L., Wu X., Yang W., York D.M. & Karplus M. (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry* **30**, 1545–1614.
cité pages 120 et 147

Bruckner S. & Boresch S. (2010). Efficiency of alchemical free energy simulations. II. improvements for thermodynamic integration. *Journal of Computational Chemistry* **32**, 1320–1333.
cité page 105

Brünger A. (1992). X-plor version 3.1, a system for X-ray crystallography and NMR. *New Haven: University Press* .
cité pages 35, 123 et 131

Buch I., Sadiq S. & De Fabritiis G. (2011). Optimized potential of mean force calculations for standard binding free energies. *Journal of Chemical Theory and Computation* **7**, 1765–1772.
cité page 166

Carlsson J., Andér M., Nervall M. & Åqvist J. (2006). Continuum solvation models in the linear interaction energy method. *Journal of Physical Chemistry B* **110**, 12034–12041.
cité page 125

Chen C., Natale D., Finn R., Huang H., Zhang J., Wu C. & Mazumder R. (2011). Representative proteomes: A stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLOS ONE* **6**, 1–9.
cité page 51

Cheng B., Montmasson M., Terradot L. & Rousselle P. (2016). Syndecans as cell surface receptors in cancer biology. A focus on their interaction with PDZ domain proteins. *Frontiers in Pharmacology* **7**.
cité page 15

Cilia E., Vuister G. & Lenaerts T. (2012). Accurate prediction of the dynamical changes within the second PDZ domain of PTP1e. *PLOS Computational Biology* **8**, e1002794.
cité page 9

Cino E., Choy W. & Karttunen M. (2012). Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations. *Journal of Chemical Theory and Computation* **8**, 2725–2740.
cité page 83

Comer J., Gumbart J., Hénin J., Lelièvre T., Pohorille A. & Chipot C. (2015). The adaptive biasing force method: Everything you always wanted to know but were afraid to a sk. *Journal of Physical Chemistry B* **119**, 1129–1151.
cité page 107

Cornell W., Cieplak P., Bayly C., Gould I., Merz K., Ferguson D., Spellmeyer D., Fox T., Caldwell J. & Kollman P. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **118**, 2309–2309.
cité pages 78, 121 et 147

Dahiyat B. & Mayo S. (1996). Protein design automation. *Protein Science* **5**, 895–903.
cité pages 27, 30 et 34

Dahiyat B.I. & Mayo S.L. (1997). De novo protein design: Fully automated sequence selection. *Science* **278**, 82–87.
cité pages 27 et 31

Darden T., York D. & Pedersen L. (1993). Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *Journal of Chemical Physics* **98**, 10089–10092.
cité pages 121, 147 et 150

de Castro E., Sigrist C., Gattiker A., Bulliard V., Langendijk-Genevaux P., Gasteiger E., Bairoch A. & Hulo N. (2006). Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic Acids Research* **34**, W362–W365.
cité page 15

de Leeuw S., Perram J. & Smith E. (1980). Simulation of electrostatic systems in periodic boundary conditions. i. lattice sums and dielectric constants. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **373**, 27–56.
cité page 150

De Maeyer M., Desmet J. & Lasters I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Folding and Design* **2**, 53–66.
cité page 24

de Ruiter A., Boresch S. & Oostenbrink C. (2013). Comparison of thermodynamic integration and bennett acceptance ratio for calculating relative protein-ligand binding free energies. *Journal of Computational Chemistry* **34**, 1024–1034.
cité page 106

Desjarlais J.R. & Handel T.M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Science* **4**, 2006–2018.
cité page 25

Desjarlais J.R. & Handel T.M. (1999). Side-chain and backbone flexibility in protein core design1. *Journal of Molecular Biology* **290**, 305–318.
cité page 25

Desmet J., De Maeyer M., Hazes B. & Lasters I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539–542.
cité page 33

Dolenc J., Oostenbrink C., Koller J. & van Gunsteren W. (2005). Molecular dynamics simulations and free energy calculations of netropsin and distamycin binding to an AAAAA DNA binding site. *Nucleic Acids Research* **33**, 725–733.
cité page 153

Doyle D., Lee A., Lewis J., Kim E., Sheng M. & MacKinnon R. (1996). Crystal structures of a complexed and peptide-free membrane protein–binding domain: Molecular basis of peptide recognition by PDZ. *Cell* **85**, 1067–1076.
cité page 6

Druart K., Bigot J., Audit E. & Simonson T. (2016). A hybrid Monte Carlo scheme for multibackbone protein design. *Journal of Chemical Theory and Computation* **12**, 6035–6048.
cité page 26

Dunbrack R. & Cohen F. (1997). Bayesian statistical analysis of protein sidechain rotamer preferences. *Protein Science* **6**, 1661–1681.
cité page 24

Dunbrack R. & Karplus M. (1993). Backbone-dependent rotamer library for proteins. application to side-chain prediction. *Journal of Molecular Biology* **230**, 543–574.
cité page 24

Durbin R., Eddy S., Krogh A. & Mitchison G. (2002). Biological sequence analysis: probabilistic models of proteins and nucleic acids. *Cambridge University Press: Cambridge, United Kingdom* .
cité page 54

Ernst A., Gfeller D., Kan Z., Seshagiri S., Kim P., Bader G. & Sidhu S. (2010). Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Molecular BioSystems* **6**, 1782–1790.
cité page 6

Ernst A., Appleton B., Ivarsson Y., Zhang Y., Gfeller D., Wiesmann C. & Sidhu S. (2014). A structural portrait of the PDZ domain family. *Journal of Molecular Biology* **426**, 3509–3519.
cité pages 6 et 7

Ferrenberg A. & Swendsen R. (1989). Optimized Monte Carlo data analysis. *Physical Review Letters* **63**, 1195–1198.
cité page 106

Finkelstein A. & Ptitsyn O. (1977). Theory of protein molecule self-organization. I. thermodynamic parameters of local secondary structures in the unfolded protein chain. *Biopolymers* **16**, 469–495.
cité page 24

Fiser A., Kinh Gian Do R. & Sali A. (2000). Modeling of loops in protein structures. *Protein Science* **9**, 1753–1773.
cité page 146

Foloppe N. & Hubbard R. (2006). Towards predictive ligand design with free-energy based computational methods? *Current Medicinal Chemistry* **13**, 3583–3608.
cité page 112

Fry D. & Vassilev L. (2005). Targeting protein-protein interactions for cancer therapy. *Journal of Molecular Medicine* **83**, 955–963.
cité page 8

Fujii N., Haresco J., Novak K., Stokoe D., Kuntz I. & Guy R. (2003). A selective irreversible inhibitor targeting a PDZ protein interaction domain. *Journal of the American Chemical Society* **125**, 12074–12075.
cité page 8

Gaillard T. & Simonson T. (2014). Pairwise decomposition of an MMGBSA energy function for computational protein design. *Journal of Computational Chemistry* **35**, 1371–1387.
cité page 31

Gainza P., Roberts E. & Donald B. (2012). Protein design using continuous rotamers. *PLOS Computational Biology* **8**, 1–15.
cité page 24

Gainza P., Roberts K., Georgiev I., Lilien R., Keedy D., Chen C., Reza F., Anderson A., Richardson D., Richardson J. & Donald B. (2013). Osprey: Protein design with ensembles, flexibility and provable algorithms. *Methods in Enzymology* **523**, 87–107.
cité page 34

Gallardo R., Ivarsson Y., Schymkowitz J., Rousseau F. & Zimmermann P. (2010). Structural diversity of PDZ-lipid interactions. *ChemBioChem* **11**, 456–467.
cité page 6

Gee S., Sekely S., Lombardo C., Kurakin A., Froehner S. & Kay B. (1998). Cyclic peptides as non-carboxyl-terminal ligands of syntrophin PDZ domains. *Journal of Biological Chemistry* **273**, 21980–21987.
cité page 5

Genheden S. & Ryde U. (2012). Comparison of end-point continuum-solvation methods for the calculation of protein–ligand binding free energies. *Proteins: Structure, Function, and Bioinformatics* **80**, 1326–1342.
cité pages 111 et 112

Genheden S. & Ryde U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* **10**, 449–461.
cité page 99

Georgiev I., Keedy D., Richardson J., Richardson D. & Donald B. (2008). Algorithm for backrub motions in protein design. *Bioinformatics* **24**, i196–i204.
cité page 26

Gilson M. & Zhou H. (2007). Calculation of protein-ligand binding affinities. *Annual review of biophysics and biomolecular structure* **36**, 21–42.
cité pages 101, 145 et 158

Gilson M., Sharp K. & Honig B. (1988). Calculating the electrostatic potential of molecules in solution: Method and error assessment. *Journal of Computational Chemistry* **9**, 327–335.
cité page 123

Gohlke H. & Case D. (2004). Converging free energy estimates: Mm-pb(gb)sa studies on the protein–protein complex ras–raf. *Journal of Computational Chemistry* **25**, 238–250.
cité page 111

Goldstein R. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophysical Journal* **66**, 1335–1340.
cité page 33

Gouda H., Kuntz I., Case D. & Kollman P. (2002). Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers* **68**, 16–34.
cité pages 112 et 114

Gough J., Karplus K., Hughey R. & Chothia C. (2001). Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure1. *Journal of Molecular Biology* **313**, 903–919.
cité page 49

Graffner-Nordberg M., Kolmodin K., Åqvist J., Queener S. & Hallberg A. (2001). Design, synthesis, computational prediction, and biological evaluation of ester soft drugs as inhibitors of dihydrofolate reductase from pneumocystis carinii. *Journal of Medicinal Chemistry* **44**, 2391–2402.
cité page 110

Gumbart J., Roux B. & Chipot C. (2013). Standard binding free energies from computer simulations: What is the best strategy? *Journal of Chemical Theory and Computation* **9**, 794–802.
cité pages 107, 166 et 168

Gutierrez-de Teran H. & Åqvist J. (2011). Linear interaction energy: Method and applications in drug design. In *Methods in Molecular Biology*, 305–323 (Springer New York).
cité page 110

Habets G., Scholtes E., Zuydgeest D., van der Kammen R., Stam J., Berns A. & Collard J. (1994). Identification of an invasion-inducing gene, Tiam-1, that encodes a protein with homology to GDP-GTP exchangers for Rho-like proteins. *Cell* **77**, 537–549.
cité page 13

Hall A. (1998). Rho GTPases and the actin cytoskeleton. *Science* **279**, 509–514.
cité page 14

Harbury P.B. Plecs J., Tidor B. Alber T. & Kim P. (1998). High-resolution protein design with backbone freedom. *Science* **282**, 1462–1467.
cité page 25

Harbury P., Tidor B. & P.S. K. (1995). Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc National Academy Sciences of U.S.A* **92**, 8408–8412.
cité page 25

Hawkins G., Cramer C. & Truhlar D. (1995). Pairwise solute descreening of solute charges from a dielectric medium. *Chemical Physics Letters* **246**, 122–129.
cité pages 123 et 131

Hillier B. (1999). Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. *Science* **284**, 812–815.
cité page 5

Holland J. (1975). Adaptation in natural and artificial systems.
cité page 33

Hom G. & Mayo S. (2005). A search algorithm for fixed-composition protein design. *Journal of Computational Chemistry* **27**, 375–378.
cité page 34

Hoover W. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **31**, 1695–1697.
cité pages 78, 121 et 147

Hou T., Guo S. & Xu X. (2002). Predictions of binding of a diverse set of ligands to gelatinase-a by a combination of molecular dynamics and continuum solvent models. *Journal of Physical Chemistry B* **106**, 5527–5535.
cité pages 112 et 114

Hou T., Chen K., McLaughlin W., Lu B. & Wang W. (2006). Computational analysis and prediction of the binding motif and protein interacting partners of the abl SH3 domain. *PLOS Computational Biology* **2**, e1.
cité pages 112 et 114

Hou T., Wang J., Li Y. & Wang W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1.The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling* **51**, 69–82.
cité pages 112 et 114

Hu G., Ma A. & Wang J. (2017). Ligand selectivity mechanism and conformational changes in guanine riboswitch by molecular dynamics simulations and free energy calculations. *Journal of Chemical Information and Modeling* **57**, 918–928.
cité page 114

Huang D., Lüthi U., Kolb P., Cecchini M., Barberis A. & Caflisch A. (2006). In silico discovery of β-secretase inhibitors. *Journal of the American Chemical Society* **128**, 5436–5443.
cité page 128

Hui S., Xing X. & Bader G. (2013). Predicting PDZ domain mediated protein interactions from structure. *BMC Bioinformatics* **14**, 27.
cité page 10

Humphrey W., Dalke A. & Schulten K. (1996). VMD - visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38.
cité page 151

Hurley J.H., Baase W.A. & Matthews B.W. (1992). Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *Journal of Molecular Biology* **224**, 1143–1159.
cité page 25

Im W., Beglov D. & Roux B. (1998). Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. *Computer Physics Communications* **111**, 59–75.
cité page 123

Ioannidis H., Drakopoulos A., Tzitzoglaki C., Homeyer N., Kolarov F., Gkeka P., Freudenberger K., Liolios C., Gauglitz G., Cournia Z., Gohlke H. & Kolocouris A. (2016). Alchemical free energy calculations and isothermal titration calorimetry measurements of aminoadamantanes bound to the closed state of influenza A/M2TM. *Journal of Chemical Information and Modeling* **56**, 862–876.
cité page 146

Ivarsson Y. (2012). Plasticity of PDZ domains in ligand recognition and signaling. *FEBS Letters* **586**, 2638–2647.
cité page 5

Jain A., Dubes R. & Chen C. (1987). Bootstrap techniques for error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**, 628–633.
cité page 149

Janin J., Wodak S., Levitt M. & Maigret B. (1978). Conformation of amino acid side-chains in proteins. *Journal of Molecular Biology* **125**, 357–386.
cité page 24

Jiang W. & Roux B. (2010). Free energy perturbation hamiltonian replica-exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. *Journal of Chemical Theory and Computation* **6**, 2559–2565.
cité page 168

Jiang W., Hodoscek M. & Roux B. (2009). Computation of absolute hydration and binding free energy with free energy perturbation distributed replica-exchange molecular dynamics. *Journal of Chemical Theory and Computation* **5**, 2583–2588.
cité page 168

Jorgensen W. & Ravimohan C. (1985). Monte Carlo simulation of differences in free energies of hydration. *Journal of Chemical Physics* **83**, 3050–3054.
cité page 103

Jorgensen W., Chandrasekhar J., Madura J., Impey R. & Klein M. (1983). Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **79**, 926–935.
cité pages 120 et 147

Joshi M., Vargas C., Boisguerin P., Diehl A., Krause G., Schmieder P., Moelling K., Hagen V., Schade M. & Oschkinat H. (2006). Discovery of low-molecular-weight ligands for the AF6 PDZ domain. *Angewandte Chemie International Edition* **45**, 3790–3795.
cité page 8

Kabsch W. & Sander C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637.
cité page 80

Kajsa B. Ljungberg K., Marelius J., Musil D., Svensson P., Norden B. & Åqvist J. (2001). Computational modelling of inhibitor binding to human thrombin. *European Journal of Pharmaceutical Sciences* **12**, 441–446.
cité page 110

Kaufmann K., Shen N., Mizoue L. & Meiler J. (2011). A physical model for PDZ-domain/peptide interactions. *Journal of Molecular Modeling* **17**, 315–324.
cité page 11

Kia A. & Darve E. (2013). The accuracy of the CHARMM22/CMAP and AMBER ff99sb force fields for modelling the antimicrobial peptide cecropin p1. *Molecular Simulation* **39**, 922–936.
cité page 83

Kim E. & Sheng M. (2004). PDZ domain proteins of synapses. *Nature Reviews Neuroscience* **5**, 771–781.
cité page 4

Kirkpatrick S., Gelatt C. & Vecchi M. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.
cité page 37

Kjellgren E., Skytte Glue O., Reinholdt P., Egeskov Meyer J., Kongsted J. & Poongavanam V. (2015). A comparative study of binding affinities for 6, 7-dimethoxy-4-pyrrolidylquinazolines as phosphodiesterase 10A inhibitors using the linear interaction energy method. *Journal of Molecular Graphics and Modelling* **61**, 44–52.
cité page 110

Koehl P. & Delarue M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *Journal of Molecular Biology* **239**, 249–275.
cité page 34

Koehl P. & Levitt M. (1999). De novo protein design. I. In search of stability and specificity. *Journal of Molecular Biology* .
cité page 27

Koller A., Schwalbe H. & Gohlke H. (2008). Starting structure dependence of NMR order parameters derived from MD simulations: Implications for judging force-field quality. *Biophysical Journal* **95**, L04–L06.
cité page 88

Kollman P. (1993). Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews* **93**, 2395–2417.
cité page 145

Kollman P., Massova I., Reyes C., Kuhn B., Huo S., Chong L., Lee M., Lee T., Duan Y., Wang W., Donini O., Cieplak P., Srinivasan J., Case D. & Cheatham T. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research* **33**, 889–897.
cité page 110

Kono H. & Doi J. (1994). Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins* **19**, 244–255.
cité page 34

Krivov G., Shapovalov M. & Dunbrack R. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics* **77**, 778–795.
cité pages 119 et 146

Kuhlman B., Dantas G., Ireton G., Varani G., Stoddard B. & D. B. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368.
cité pages 26 et 34

Kuhn B. & Kollman P. (2000). Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *Journal of Medicinal Chemistry* **43**, 3786–3791.
cité pages 112 et 114

Kumar S., Rosenberg J., Bouzida D., Swendsen R. & Kollman P. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **13**, 1011–1021.
cité page 106

Kundu K. & Backofen R. (2014). Cluster based prediction of PDZ-peptide interactions. *BMC Genomics* **15**, S5.
cité page 10

Lamb M., Tirado-Rives J. & Jorgensen W. (1999). Estimation of the binding affinities of FKBP12 inhibitors using a linear response method. *Bioorganic & Medicinal Chemistry* **7**, 851–860.
cité page 110

Lamoureux G. & Roux B. (2003). Modeling induced polarization with classical drude oscillators: Theory and molecular dynamics simulation algorithm. *Journal of Chemical Physics* **119**, 3025–3039.
cité page 169

Lamoureux G., MacKerell A. & Roux B. (2003). A simple polarizable model of water based on classical drude oscillators. *Journal of Chemical Physics* **119**, 5185–5197.
cité page 169

Landau L. & Lifshitz E. (1938). In *Statistical physics* (Clarendon Press : Oxford).
cité page 145

Landau L. & Lifshitz E. (1951). Statischeskaia fizika. *Moscow* 112–115.
cité page 104

Lauck F., Smith C.A., Friedland G.F., Humphris E.L. & Kortemme T. (2010). Rosettabackrub - a web server for flexible backbone protein structure modeling and design. *Nucleic Acids Research* **38**, W569–W575.
cité page 11

Launay G., Mendez R., Wodak S. & Simonson T. (2007). Recognizing protein–protein interfaces with empirical potentials and reduced amino acid alphabets. *BMC Bioinformatics* **8**, 270.
cité pages 44 et 55

Lazaridis T. & Karplus M. (1999). Effective energy function for proteins in solution. *Proteins: Structure, Function, and Bioinformatics* **35**, 133–152.
cité pages 37 et 132

LeBlanc B., Iwata M., Mallon A., Rupasinghe C., Goebel D., Marshall J., Spaller M. & Saab C. (2010). A cyclic peptide targeted against PSD-95 blocks central sensitization and attenuates thermal hyperalgesia. *Neuroscience* **167**, 490–500.
cité page 8

Lee C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *Journal of Molecular Biology* **236**, 918–939.
cité page 34

Lepsik M., Kriz Z. & Havlas Z. (2004). Efficiency of a second-generation HIV-1 protease inhibitor studied by molecular dynamics and absolute binding free energy calculations. *Proteins: Structure, Function, and Bioinformatics* **57**, 279–293.
cité page 111

Lin K., Simossis V., Taylor W. & Heringa J. (2004). A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* **21**, 152–159.
cité page 76

Lindorff-Larsen K., Maragakis P., Piana S., Eastwood M., Dror R. & Shaw D. (2012). Systematic validation of protein force fields against experimental data. *PLOS ONE* **7**, e32131.
cité page 83

Liu P., Dehez F., Cai W. & Chipot C. (2012). A toolkit for the analysis of free-energy perturbation calculations. *Journal of Chemical Theory and Computation* **8**, 2606–2616.
cité page 148

Liu X., Shepherd T., Murray A., Xu Z. & Fuentes E. (2013). The structure of the Tiam1 PDZ domain phospho-Syndecan1 complex reveals a ligand conformation that modulates protein dynamics. *Structure* **21**, 342–354.
cité pages 15, 16, 17, 18, 72, 146, 160 et 162

Liu X., Speckhard D., Shepherd T., Sun Y., Hengel S., Yu L., Fowler A., Gakhar L. & Fuentes E. (2016). Distinct roles for conformational dynamics in protein-ligand interactions. *Structure* **24**, 2053–2066.
cité pages 16, 17, 18, 62, 79 et 146

Long J., Wei Z., Feng W., Yu C., Zhao Y. & Zhang M. (2008). Supramodular nature of GRIP1 revealed by the structure of its PDZ12 tandem in complex with the carboxyl tail of fras1. *Journal of Molecular Biology* **375**, 1457–1468.
cité page 5

Lopes A., Alexandrov A., Bathelt C., Archontis G. & Simonson T. (2007). Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins: Structure, Function, and Bioinformatics* **67**, 853–867.
cité page 123

Lovell S., Word J., Richardson J. & Richardson D. (2000). The penultimate rotamer library. *Proteins* **40**, 389–408.
cité page 24

Luck K., Charbonnier S. & Travé G. (2012). The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. *FEBS Letters* **586**, 2648–2661.
cité page 4

López M., Lacroix E., Ramírez-Alvarado M. & Serrano L. (2001). Computer-aided design of β-sheet peptides. *Journal of Molecular Biology* **312**, 229–246.
cité page 12

Mach P. & Koehl P. (2013). Capturing protein sequence–structure specificity using computational sequence design. *Proteins: Structure, Function, and Bioinformatics* **81**, 1556–1570.
cité page 69

Mamonova T., Zhang Q., Chandra M., Collins B., Sarfo E., Bu Z., Xiao K., Bisello A. & Friedman P. (2017) **56**, 2584–2593.
cité page 10

Marcus R. (1964). Chemical and electrochemical electron-transfer theory. *Annual Review of Physical Chemistry* **15**, 155–196.
cité page 109

Marshall S., Vizcarra C. & Mayo S. (2005). One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. *Protein Science* **14**, 1293–1304.
cité page 30

Masili C., Schumann M., Toussaint N., Kageyama J., Kohlbacher O. & Hocker B. (2012). Binding pocket optimization by computational protein design. *PLOS ONE* **7**, e52505.
cité page 34

Masuda M., Maruyama T., Ohta T., Ito A., Hayashi T., Tsukasaki K., Kamihira S., Yamaoka S., Hoshino H., Yoshida T., Watanabe T., Stanbridge E. & Murakami Y. (2010). CADM1 interacts with Tiam1 and promotes invasive phenotype of human T-cell leukemia virus type

I-transformed cells and adult T-cell leukemia cells. *Journal of Biological Chemistry* **285**, 15511–15522.
cité page 15

Mayasundari A., Ferreira A., He L., Mahindroo N., Bashford D. & Fujii N. (2008). Rational design of the first small-molecule antagonists of NHERF1/EBP50 PDZ domains. *Bioorganic & Medicinal Chemistry Letters* **18**, 942–945.
cité page 8

Melero C., Ollikainen N., Harwood I., Karpiak J. & Kortemme T. (2014). Quantification of the transferability of a designed protein specificity switch reveals extensive epistasis in molecular recognition. *Proceedings of the National Academy of Sciences* **111**, 15426–15431.
cité page 12

Metropolis N., Rosenbluth A., Rosenbluth M., Teller A. & Teller E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.
cité page 32

Michael E., Polydorides S., Simonson T. & Archontis G. (2017). Simple models for nonpolar solvation: Parameterization and testing. *Journal of Computational Chemistry* 1–11.
cité page 132

Mignon D. & Simonson T. (2016). Comparing three stochastic search algorithms for computational protein design: Monte Carlo, replica exchange Monte Carlo, and a multistart, steepest-descent heuristic. *Journal of Computational Chemistry* **37**, 1781–1793.
cité page 36

Mignon D., Panel N., Chen X., Fuentes E. & Simonson T. (2017). Computational design of the Tiam1 PDZ domain and its ligand binding. *Journal of Chemical Theory and Computation* **13**, 2271–2289.
cité page 172

Mikulskis P., Genheden S. & Ryde U. (2014). A large-scale test of free-energy simulation estimates of protein–ligand binding affinities. *Journal of Chemical Information and Modeling* **54**, 2794–2806.
cité page 158

Minard M., Kim L., Price J. & Gallick G. (2004). The role of the guanine nucleotide exchange factor Tiam1 in cellular migration, invasion, adhesion and tumor progression. *Breast Cancer Research and Treatment* **84**, 21–32.
cité page 14

Minard M., Herynk M., Collard J. & Gallick G. (2005). The guanine nucleotide exchange factor Tiam1 increases colon carcinoma growth at metastatic sites in an orthotopic nude mouse model. *Oncogene* **24**, 2568–2573.
cité page 14

Minard M., Ellis L. & Gallick G. (2006). Tiam1 regulates cell adhesion, migration and apoptosis in colon tumor cells. *Clinical & Experimental Metastasis* **23**, 301–313.
cité page 14

Mobley D. & Klimovich P. (2012). Perspective: Alchemical free energy calculations for drug discovery. *Journal of chemical physics* **137**, 230901+.
cité page 150

Mobley D., Chodera J. & Dill K. (2007a). Confine-and-release method: Obtaining correct binding free energies in the presence of protein conformational change. *Journal of Chemical Theory and Computation* **3**, 1231–1235.
cité page 168

Mobley D., Graves A., Chodera J., McReynolds A., Shoichet B. & Dill K. (2007b). Predicting absolute ligand binding free energies to a simple model site. *Journal of Molecular Biology* **371**, 1118–1134.
cité page 168

Mooers B.H., Datta D., Baase W.A., Zollars E.S., Mayo S.L. & Matthews B.W. (2003). Repacking the core of T4 lysozyme by automated design. *Journal of Molecular Biology* **332**, 741–756.
cité page 25

Murciano-Calles J., McLaughlin M., Erijman A., Hooda Y., Chakravorty N., Martinez J., Shifman J. & Sidhu S. (2014). Alteration of the C-terminal ligand specificity of the erbin PDZ domain by allosteric mutational effects. *Journal of Molecular Biology* **426**, 3500–3508.
cité page 6

Murphy L., Wallqvist A. & Levy R. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering, Design and Selection* **13**, 149.
cité page 55

Nakariyakul S., Liu Z. & Chen L. (2014). A sequence-based computational approach to predicting PDZ domain-peptide interactions. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1844**, 165–170.
cité page 10

Nosé S. (1984). A unified formulation of the constant temperature molecular dynamics methods. *Journal of Chemical Physics* **81**, 511–519.
cité pages 78, 121 et 147

Nourry C., Grant S. & Borg J. (2003). PDZ domain proteins: Plug and play! *Science Signaling* **2003**, 1–13.
cité page 7

Offer G. & Sessions R. (1995). Computer modelling of the alpha-helical coiled coil: packing of side-chains in the inner core. *Journal of Molecular Biology* **249**, 967–987.
cité page 25

Olsson M., Søndergaard C., Rostkowski M. & Jensen J. (2011). PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. *Journal of Chemical Theory and Computation* **7**, 525–537.
cité pages 75 et 120

Onufriev A., Case D. & Bashford D. (2002). Effective born radii in the generalized Born approximation: The importance of being perfect. *Journal of Computational Chemistry* **23**, 1297–1304.
cité page 123

Paliwal H. & Shirts M. (2011). A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods. *Journal of Chemical Theory and Computation* **7**, 4115–4134.
cité page 106

Panel N., Sun Y., Fuentes E. & Simonson T. (2017). A simple PB/LIE free energy function accurately predicts the peptide binding specificity of the Tiam1 PDZ domain. *Frontiers in Molecular Biosciences - Molecular Recognition* in press.
cité page 172

Park H., Bradley P., Greisen P., Liu Y., Mulligan V.K., Kim D.E., Baker D. & DiMaio F. (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of Chemical Theory and Computation* **12**, 6201–6212.
cité page 36

Patra C., Rupasinghe C., Dutta S., Bhattacharya S., Wang E., Spaller M. & Mukhopadhyay D. (2012). Chemically modified peptides targeting the PDZ domain of GIPC as a therapeutic approach for cancer. *ACS Chemical Biology* **7**, 770–779.
cité page 8

Pedersen S., Moran G., Sereikaitė V., Haugaard-Kedström L. & Strømgaard K. (2016). Importance of a conserved lys/arg residue for ligand/PDZ domain interactions as examined by protein semisynthesis. *ChemBioChem* **17**, 1936–1944.
cité page 72

Peierls R. (1933). On the theory of diamagnetism of conduction electrons. *Zeitschrift für Physiks* **80**, 763–791.
cité page 104

Phillips J., Braun R., Wang W., Gumbart J., Tajkhorshid E., Villa E., Chipot C., Skeel R., Kalé L. & Schulten K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781–1802.
cité pages 78, 121 et 147

Pollastri G. & McLysaght A. (2004). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* **21**, 1719–1720.
cité page 76

Polydorides S., Michael E., Mignon D., Druart K., Archontis G. & Simonson T. (2016). Proteus and the design of ligand binding sites. *Methods in Molecular Biology: Computational Design of Ligand Binding Proteins* **1414**, 77–97.
cité page 34

Ponder J. & Richards F. (1987). Tertiary templates for proteins. *Journal of Molecular Biology* **193**, 775–791.
cité page 24

Price D. & Jorgensen W. (2000). Computational binding studies of human pp60c-src SH2 domain with a series of nonpeptide, phosphophenyl-containing ligands. *Bioorganic & Medicinal Chemistry Letters* **10**, 2067–2070.
cité page 153

Raha K., Wollacott A., Italia M. & Desjarlais J. (2000). Prediction of amino acid sequence from structure. *Protein Science* **9**, 1106–1119.
cité page 30

Raimondi F., Felline A., Seeber M., Mariani S. & Fanelli F. (2013). A mixed protein structure network and elastic network model approach to predict the structural communication in biomolecular systems: The PDZ2 domain from tyrosine phosphatase 1E as a case study. *Journal of Chemical Theory and Computation* **9**, 2504–2518.
cité page 9

Rastelli G., Del Rio A., Degliesposti G. & Sgobba M. (2010). Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *Journal of Computational Chemistry* **31**, 797–810.
cité pages 112 et 114

Reiland J., Ott V., Lebakken C., Yeaman C., McCarthy J. & Rapraeger A. (1996). Pervanadate activation of intracellular kinases leads to tyrosine phosphorylation and shedding of syndecan-1. *Biochemical Journal* **319**, 39–47.
cité page 162

Reina J., Lacroix E., Hobson S., Fernandez-Ballester G., Rybin V., Schwab M., Serrano L. & Gonzalez C. (2002). Computer-aided design of a PDZ domain to recognize new target sequences. *Nature Structural Biology* .
cité page 12

Roberts K., Cushing P., Boisguerin P., Madden D. & Donald B. (2012). Computational design of a PDZ domain peptide inhibitor that rescues CFTR activity. *PLOS Computational Biology* **8**, e1002477.
cité page 12

Roothaan C.C.J. (1951). New developments in molecular orbital theory. *Rev. Mod. Phys.* **23**, 69–89.
cité page 120

Rousselle P. & Beck K. (2013). Laminin 332 processing impacts cellular behavior. *Cell Adhesion & Migration* **7**, 122–134.
cité page 14

Ryckaert J., Ciccotti G. & Berendsen H. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327–341.
cité page 78

Sali A. & Blundell T. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* **234**, 779–815.
cité page 146

Saunders C. & Baker D. (2005). Recapitulation of protein family divergence using flexible backbone protein design. *Journal of Molecular Biology* **346**, 631–644.
cité page 26

Schmidt Am Busch M., Lopes A., Mignon D. & Simonson T. (2008). Computational protein design: software implementation, parameter optimization, and performance of a simple model. *Journal of Computational Chemistry* **29**, 1092–1102.
cité page 34

Schrauber H., Eisenhaber F. & Argos P. (1993). Rotamers: to be or not to be? an analysis of amino acid sidechain conformations in globular proteins. *Journal of Molecular Biology* **23**, 592–612.
cité page 24

Schutz C. & Warshel A. (2001). What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Structure, Function, and Bioinformatics* **44**, 400–417.
cité page 111

Sensoy O. & Weinstein H. (2015). A mechanistic role of helix 8 in GPCRs: Computational modeling of the dopamine d2 receptor interaction with the GIPC1–PDZ-domain. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1848**, 976–983.
cité page 9

Shapovalov M. & Dunbrack R. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858.
cité page 36

Shepherd T., Klaus S., Liu X., Ramaswamy S., DeMali K. & Fuentes E. (2010). The Tiam1 PDZ domain couples to Syndecan1 and promotes cell–matrix adhesion. *Journal of Molecular Biology* **398**, 730–746.
cité pages 15, 16, 17, 18, 72 et 146

Shepherd T., Hard R., Murray A., Pei D. & Fuentes E. (2011). Distinct ligand specificity of the Tiam1 and Tiam2 PDZ domains. *Biochemistry* **50**, 1296–1308.
cité pages 15, 16, 17, 18, 60, 65, 72, 118, 138 et 154

Shirts M. & Chodera J. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics* **129**, 124105.
cité page 106

Shirts M. & Pande V. (2005). Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration. *Journal of Chemical Physics* **122**, 144107.
cité page 105

Shoup D. & Szabo A. (1982). Role of diffusion in ligand binding to macromolecules and cell-bound receptors. *Biophysical Journal* **40**, 33–39.
cité page 106

Simonson T. (2001). Free energy calculations. In *Computational Biochemistry and Biophysics*, 169–197 (CRC Press).
cité page 100

Simonson T. (2013). Protein: Ligand recognition: Simple models for electrostatic effects. *Current Pharmaceutical Design* **19**, 4241–4256.
cité page 135

Simonson T. (2015). The physical basis of ligand binding. In C.N. Cavasotto, ed., *In Silico Drug Discovery and Design.*, 3–43 (Edited by Cavasotto, C.N. - CRC Press).
cité page 99

Simonson T., Archontis G. & Karplus M. (2002). Free energy simulations come of age: Protein-ligand recognition. *Accounts of Chemical Research* **35**, 430–437.
cité page 169

Simonson T., Carlsson J. & Case D. (2004). Proton binding to proteins: pKa calculations with explicit and implicit solvent models. *Journal of the American Chemical Society* **126**, 4167–4180.
cité page 166

Simonson T., Gaillard T., Mignon D., Schmidt am Busch M., Lopes A., Amara N., Polydorides S., Sedano A., Druart K. & Archontis G. (2013). Computational protein design: The proteus software and selected applications. *Journal of Computational Chemistry* **37**, 2472–2484.
cité page 34

Singh U. & Kollman P. (1984). An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **5**, 129–145.
cité page 120

Smith C. & Kortemme T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of Molecular Biology* **380**, 742–756.
cité page 26

Smith C. & Kortemme T. (2010). Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *Journal of Molecular Biology* **402**, 460–474.
cité page 10

Smith M., Rao J., Segelken E. & Cruz L. (2015). Force-field induced bias in the structure of $a_{\beta 21-30}$: A comparison of OPLS, AMBER, CHARMM, and GROMOS force fields. *Journal of Chemical Information and Modeling* **55**, 2587–2595.
cité page 83

Smith R., Jorgensen W., Tirado-Rives J., Lamb M., Janssen P., Michejda C. & Kroeger Smith M. (1998). Prediction of binding affinities for TIBO inhibitors of HIV-1 reverse transcriptase using Monte Carlo simulations in a linear response method. *Journal of Medicinal Chemistry* **41**, 5272–5286.
cité page 110

Songyang Z., Fanning A., Fu C., Xu J., Marfatia S.M., Chishti A.H., Crompton A., Chan A.C., Anderson J.M. & Cantley L.C. (1997). Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science* **275**, 73–77.
cité pages 6, 7, 15 et 146

Sonnhammer E., Eddy S. & Durbin R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics* **28**, 405–420.
cité page 51

Spiegel I., Salomon D., Erne B., Schaeren-Wiemers N. & Peles E. (2002). Caspr3 and Caspr4, two novel members of the Caspr family are expressed in the nervous system and interact with PDZ domains. *Molecular and Cellular Neuroscience* **20**, 283–297.
cité page 15

Spiliotopoulos D., Spitaleri A. & Musco G. (2012). Exploring PHD fingers and H3K4me0 interactions with molecular dynamics simulations and binding free energy calculations: AIRE-PHD1, a comparative study. *PLOS ONE* **7**, 1–13.
cité pages 112 et 114

Srinivasan J., Cheatham T., Cieplak P., Kollman P. & Case D. (1998). Continuum solvent studies of the stability of dna, rna, and phosphoramidate−DNA helices. *Journal of the American Chemical Society* **120**, 9401–9409.
cité page 110

Stiffler M., Chen J., Grantcharova V., Lei Y., Fuchs D., Allen J., Zaslavskaia L. & MacBeath G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. *Science* **317**, 364–369.
cité pages 8 et 146

Stoica I., Sadiq S. & Coveney P. (2008). Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. *Journal of the American Chemical Society* **130**, 2639–2648.
cité pages 112 et 114

Straatsma T. & McCammon J. (1991). Theoretical calculations of relative affinities of binding. *Methods in Enzymology* **202**, 497–511.
cité page 103

Straatsma T. & McCammon J. (1992). Computational alchemy. *Annual Review of Physical Chemistry* **43**, 407–435.
cité pages 101 et 145

Street A. & Mayo S. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Folding and Design* **3**, 253–258.
cité page 32

Stricker N., Christopherson K., Yi B., Schatz P., Raab R., Dawes G., Bassett D., Bredt D. & Li M. (1997). PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. *Nature Biotechnology* **15**, 336–342.
cité page 7

Su A. & Mayo S.L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Science* **6**, 1701–1707.
cité page 25

Suárez M. & Jaramillo A. (2009). Challenges in the computational design of proteins. *Journal of The Royal Society Interface* **6**, S477–S491.
cité page 27

Sulka B., Lortat-Jacob H., Terreux R., Letourneur F. & Rousselle P. (2009). Tyrosine dephosphorylation of the Syndecan-1 PDZ binding domain regulates syntenin-1 recruitment. *Journal of Biological Chemistry* **284**, 10659–10671.
cité pages 14 et 162

Sunhwan J., Taehoon K., Vidyashankara I. & Wonpil I. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **29**, 1859–1865.
cité pages 120 et 147

Swanson J., Henchman R. & McCammon J. (2004). Revisiting free energy calculations: A theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophysical Journal* **86**, 67–74.
cité pages 111 et 135

Swanson J., Adcock S. & McCammon J. (2005). Optimized radii for Poisson−Boltzmann calculations with the AMBER force field. *Journal of Chemical Theory and Computation* **1**, 484–493.
cité page 123

Tembe B. & McCammon J. (1984). Ligand-receptor interactions. *Computers & Chemistry* **8**, 281–283.
cité pages 103 et 146

Thorsen T., Madsen K., Rebola N., Rathje M., Anggono R., Bach A., Moreira I., Stuhr-Hansen N., Dyhring T., Peters D., Beuming T., Huganir R., Weinstein H., Mulle C., Stromgaard K., Ronn L. & Gether U. (2009). Identification of a small-molecule inhibitor of the PICK1 PDZ domain that inhibits hippocampal LTP and LTD. *Proceedings of the National Academy of Sciences* **107**, 413–418.
cité page 8

Tian F., Lv L., Zhou P. & Yang L. (2011). Characterization of PDZ domain-peptide interactions using an integrated protocol of QM/MM, PB/SA, and CFEA analyses. *Journal of Computer-Aided Molecular Design* **25**, 947–958.
cité page 10

Tonikian R., Zhang Y., Sazinsky S., Currell B., Yeh J., Reva B., Held H., Appleton B., Evangelista M., Wu Y., Xin X., Chan A.C., Seshagiri S., Lasky L.A., Sander C., Boone C., Bader G.D. & Sidhu S. (2008). A specificity map for the PDZ domain family. *PLOS Biology* **6**, 1–17.
cité pages 6, 7, 15 et 146

Towse C.L., Rysavy S., Vulovic I. & Daggett V. (2016). New dynamic rotamer libraries: Data-driven analysis of side-chain conformational propensities. *Structure* **24**, 187–199.
cité page 24

Tuffery P., Etchebest C., Hazout S. & Lavery R. (1991). A new approach to the rapid determination of protein side chain conformations. *Journal of Biomolecular Structure and Dynamic* **8**, 1267–1289.
cité page 24

Tuffery P., Etchebest C. & Hazout S. (1997). Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Engineering Design and Selection* **10**, 361–372.
cité page 35

Udugamasooriya D., Sharma S. & Spaller M. (2008). A chemical library approach to organic-modified peptide ligands for PDZ domain proteins: A synthetic, thermodynamic and structural investigation. *ChemBioChem* **9**, 1587–1589.
cité page 8

Vaccaro P. & Dente L. (2002). PDZ domains: troubles in classification. *FEBS Letters* **512**, 345–346.
cité page 7

Valleau J. & Card D. (1972). Monte Carlo estimation of the free energy by multistage sampling. *Journal of Chemical Physics* **57**, 5457–5462.
cité page 101

Venken T., Krnavek D., Münch J., Kirchhoff F., Henklein P., De Maeyer M. & Voet A. (2011). An optimized MM/PBSA virtual screening approach applied to an HIV-1 gp41 fusion peptide inhibitor. *Proteins: Structure, Function, and Bioinformatics* **79**, 3221–3235.
cité pages 112 et 114

Villa A., Zangi R., Pieffet G. & Mark A. (2003). Sampling and convergence in free energy calculations of protein-ligand interactions: The binding of triphenoxypyridine derivatives to factor Xa and trypsin. *Journal of Computer-Aided Molecular Design* **17**, 673–686.
cité page 153

Vizcarra C., Zhang N., Marshall S., Wingreen N., Zeng C. & Mayo S. (2008). An improved pairwise decomposable finite-difference Poisson–Boltzmann method for computational protein design. *Journal of Computational Chemistry* **29**, 1153–1162.
cité page 30

Wall I., Leach A., Salt D., Ford M. & Essex J. (1999). Binding constants of neuraminidase inhibitors: An investigation of the linear interaction energy method. *Journal of Medicinal Chemistry* **42**, 5142–5152.
cité page 110

Wan S., Coveney P. & Flower D. (2005). Molecular basis of peptide recognition by the TCR: Affinity differences calculated using large scale computing. *Journal of Immunology* **175**, 1715–1723.
cité page 146

Wan S., Knapp B., Wright D., Deane C. & Coveney P. (2015). Rapid, precise, and reproducible prediction of peptide–MHC binding affinities from molecular dynamics that correlate well with experiment. *Journal of Chemical Theory and Computation* **11**, 3346–3356.
cité page 112

Wang C., Pan L., Chen J. & Zhang M. (2010). Extensions of PDZ domains as important structural and functional elements. *Protein & Cell* **1**, 737–751.
cité page 5

Wang L., Deng Y., Knight J., Wu Y., Kim B., Sherman W., Shelley J., Lin T. & Abel R. (2013). Modeling local structural rearrangements using FEP/REST: Application to relative binding affinity predictions of CDK2 inhibitors. *Journal of Chemical Theory and Computation* **9**, 1282–1293.
cité page 150

Wang W. & Kollman P. (2000). Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. *Journal of Molecular Biology* **303**, 567–582.
cité pages 112 et 114

Wang W., Wang J. & Kollman P. (1999). What determines the van der waals coefficient β in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? *Proteins: Structure, Function, and Bioinformatics* **34**, 395–402.
cité page 109

Warshel A., Sussman F. & King G. (1986). Free energy of charges in solvated proteins: microscopic calculations using a reversible charging process. *Biochemistry* **25**, 8368–8372.
cité page 100

Welch B., VanDemark A., Heroux A., Hill C. & Kay M. (2007). Potent D-peptide inhibitors of HIV-1 entry. *Proceedings of the National Academy of Sciences* **104**, 16828–16833.
cité page 142

Wells J. & McClendon C. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **450**, 1001–1009.
cité page 8

Wernisch L., Hery S. & Wodak S. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. *Journal of Molecular Biology* **301**, 713–736.
cité pages 27 et 33

Wesson L. & Eisenberg D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science* **1**, 227–235.
cité page 30

Wilson D., Madera M., Vogel C., Chothia C. & Gough J. (2007). The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Research* **35**, D308–D313.
cité page 49

Wong C. & McCammon J. (1986). Dynamics and design of enzymes and inhibitors. *Journal of the American Chemical Society* **108**, 3830–3832.
cité page 100

Wong H., Bourdelas A., Krauss A., Lee H., Shao Y., Wu D., Mlodzik M., Shi D. & Zheng J. (2003). Direct binding of the PDZ domain of dishevelled to a conserved internal sequence in the C-terminal region of frizzled. *Molecular Cell* **12**, 1251–1260.
cité page 5

Woo H. & Roux B. (2005). Calculation of absolute protein-ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences* **102**, 6825–6830.
cité pages 107, 108, 146, 166 et 169

Xiao X., He Q., Yu L., Wang S., Li Y., Yang H., Zhang A., Ma X., Peng Y. & Chen B. (2017). Structure-based optimization of salt-bridge network across the complex interface of PTPN4 PDZ domain with its peptide ligands in neuroglioma. *Computational Biology and Chemistry* **66**, 63–68.
cité page 12

Yamaguchi H., Kodama H., Osada S., Kato F., Jelokhani-niaraki M. & Kondo M. (2003). Effect of α,α-dialkyl amino acids on the protease resistance of peptides. *Bioscience, Biotechnology, and Biochemistry* **67**, 2269–2272.
cité page 119

Yang C., Sun H., Chen J., Nikolovska-Coleska Z. & Wang S. (2009). Importance of ligand reorganization free energy in protein-ligand binding-affinity prediction. *Journal of the American Chemical Society* **131**, 13709–13721.
cité page 111

Ye F. & Zhang M. (2013). Structures and target recognition modes of PDZ domains: recurring themes and emerging pictures. *Biochemical Journal* **455**, 1–14.
cité pages 4 et 5

Yen-Lin L., Alexey A., Thomas S. & Benoît R. (2014). An overview of electrostatic free energy computations for solutions and proteins. *Journal of Chemical Theory and Computation* **10**, 2690–2709.
cité page 150

Zacharias M., Straatsma T. & McCammon J. (1994). Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *Journal of Chemical Physics* **30**, 9025–9031.
cité page 148

Zheng F., Jewell H., Fitzpatrick J., Zhang J., Mierke D. & Grigoryan G. (2015). Computational design of selective peptides to discriminate between similar PDZ domains in an oncogenic pathway. *Journal of Molecular Biology* **427**, 491–510.
cité page 12

Zwanzig R. (1954). High-temperature equation of state by a perturbation method. I. Nonpolar gases. *Journal of Chemical Physics* **22**, 1420–1426.
cité pages 104 et 145

université PARIS-SACLAY
ÉCOLE DOCTORALE
INTERFACES
Approches Interdisciplinaires : Fondements, Applications et Innovation

Titre : Étude computationnelle du domaine PDZ de Tiam1

Mots clefs : Dessin de protéine, calcul d'énergie libre, simulation de dynamique moléculaire, interactions protéine-peptide, Tiam1, PDZ

Résumé : Les interactions protéine-protéine sont souvent contrôlées par de petits domaines protéiques qui régulent les chemins de signalisation au sein des cellules eucaryotes. Les domaines PDZ sont parmi les domaines les plus répandus et les plus étudiés. Ils reconnaissent spécifiquement les 4 à 10 acides aminés C-terminaux de leurs partenaires. Tiam1 est un facteur d'échange de GTP de la protéine Rac1 qui contrôle la migration et la prolifération cellulaire et dont le domaine PDZ lie les protéines Syndecan-1 (Sdc1), Caspr4 et Neurexine. Des petits peptides ou des molécules peptidomimétiques peuvent potentiellement inhiber ou moduler son activité et être utilisés à des fins thérapeutiques. Nous avons appliqué des approches de dessin computationnel de protéine (CPD) et de calcul d'énergie libre par simulations dynamique moléculaire (DM) pour comprendre et modifier sa spécificité. Le CPD utilise un modèle structural et une fonction d'énergie pour explorer l'espace des séquences et des structures et identifier des variants protéiques ou peptidiques stables et fonctionnels. Nous avons utilisé le programme de CPD Proteus, développé au laboratoire, pour redessiner entièrement le domaine PDZ de Tiam1. Les séquences générées sont similaires à celles des domaines PDZ naturels, avec des scores de similarité et de reconnaissance de pli comparables au programme Rosetta, un outil de CPD très utilisé. Des séquences contenant environ 60 positions mutées sur 90, ont été testées par simulations de DM et des mesures biophysiques. Quatre des cinq séquences testées expérimentalement (par nos collaborateurs) montrent un dépliement réversible autour de 50°C. Proteus a également déterminer correctement la spécificité de la liaison de quelques variants protéiques et peptidiques. Pour étudier plus finement la spécificité, nous avons paramétré un modèle d'énergie libre semi-empirique de Poisson-Boltzmann ayant la forme d'une énergie linéaire d'interaction, ou PB/LIE, appliqué à des conformations issues de simulations de DM en solvant explicite de complexes PDZ:peptide. Avec trois paramètres ajustables, le modèle reproduit correctement les affinités expérimentales de 41 variants, avec une erreur moyenne absolue de 0,4 kcal/mol, et donne des prédictions pour 10 nouveaux variants. Le modèle PB/LIE a ensuite comparé à la méthode non-empirique de calcul d'énergie libre par simulations alchimiques, qui n'a pas de paramètre ajustable et qui prédit correctement l'affinité de 12 complexes Tiam1:peptide. Ces outils et les résultats obtenus devraient nous permettre d'identifier des peptides inhibiteurs et auront d'importantes retombées pour l'ingénierie des interactions PDZ:peptide.

Title : Computational study of the Tiam1 PDZ domain

Keywords : Computational Protein Design, free energy calculation, molecular dynamics simulation, protein-peptide interactions, Tiam1, PDZ

Abstract : Small protein domains often direct protein-protein interactions and regulate eukaryotic signalling pathways. PDZ domains are among the most widespread and best-studied. They specifically recognize the 4-10 C-terminal amino acids of target proteins. Tiam1 is a Rac GTP exchange factor that helps control cell migration and proliferation and whose PDZ domain binds the proteins syndecan-1 (Sdc1), Caspr4, and Neurexin. Short peptides and peptidomimetics can potentially inhibit or modulate its action and act as bioreagents or therapeutics. We used computational protein design (CPD) and molecular dynamics (MD) free energy simulations to understand and engineer its peptide specificity. CPD uses a structural model and an energy function to explore the space of sequences and structures and identify stable and functional protein or peptide variants. We used our in-house Proteus CPD package to completely redesign the Tiam1 PDZ domain. The designed sequences were similar to natural PDZ domains, with similarity and fold recognition scores comparable to the widely-used Rosetta CPD package. Selected sequences, containing around 60 mutated positions out of 90, were tested by microsecond MD simulations and biophysical experiments. Four of five sequences tested experimentally (by our collaborators) displayed reversible unfolding around 50°. Proteus also accurately scored the binding specificity of several protein and peptide variants. As a more refined model for specificity, we parameterized a semi-empirical free energy model of the Poisson-Boltzmann Linear Interaction Energy or PB/LIE form, which scores conformations extracted from explicit solvent MD simulations of PDZ:peptide complexes. With three adjustable parameters, the model accurately reproduced the experimental binding affinities of 41 variants, with a mean unsigned error of just 0.4 kcal/mol, and gave predictions for 10 new variants. The PB/LIE model was tested further by comparing to non-empirical, alchemical, MD free energy simulations, which have no adjustable parameters and were found to give chemical accuracy for 12 Tiam1:peptide complexes. The tools and insights obtained should help discover new tight binding peptides or peptidomimetics and have broad implications for engineering PDZ:peptide interactions.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France