Evolution dirigée de deux aminoacyl-ARNt synthétases : Mise en place et applications d'une méthode de 'protein design'.

Anne Lopes

ÉCOLE POLYTECHNIQUE

THESE
pour obtenir le titre de
DOCTEUR DE L'ECOLE POLYTECHNIQUE
Discipline : Biologie

Présentée et soutenue publiquement par
Anne LOPES
le 30 janvier 2008

Evolution dirigée *in silico*: mise en place d'une méthode théorique de *protein design* et application à deux aminoacyl-ARNt synthétases

Jury composé de

Dr. Marc	Delarue	Rapporteur
Pr. Gilbert	Deléage	Rapporteur
Dr. Jacques	Chomilier	Président
Dr. Raphaël	Guerois	Examinateur
Pr. Martin	Karplus	Examinateur
Pr. Thomas	Simonson	Directeur de thèse
Dr. Pierre	Tufféry	Examinateur

RESUME

La conception des protéines ou 'protein design' a pour but de développer des protéines possédant de nouvelles caractéristiques structurales et/ou fonctionnelles. Le principe consiste à identifier parmi toutes les séquences compatibles avec le repliement d'intérêt, celles qui vont conférer à la protéine, la fonction désirée. La procédure générale est réalisée en deux étapes. La première consiste à calculer une matrice d'énergie contenant les énergies d'interactions entre toutes les paires de résidus de la protéine en autorisant successivement tous les types d'acides aminés dans toutes leurs conformations possibles. La seconde étape, ou 'phase d'optimisation', consiste à explorer simultanément l'espace des séquences et des conformations afin de déterminer la combinaison optimale d'acides aminés étant donné le repliement de départ. Ensuite, différents filtres peuvent être appliqués pour sélectionner les séquences fonctionnelles (étant donné le repliement d'intérêt) des non fonctionnelles.

La première étape a consisté au développement de la procédure de 'protein design', en particulier, à la mise en place et à l'optimisation de la fonction d'énergie ainsi qu'à l'implémentation de l'algorithme d'optimisation. Nous avons montré que notre procédure est robuste puisqu'elle a fait ses preuves dans des applications très diverses telles que la prédiction de l'orientation des chaînes latérales, la prédiction des changements de stabilité ou d'affinité associés à des mutations ponctuelles, ou encore la production de séquences de type natif pour un jeu de protéines globulaires. Pour l'ensemble de ces applications, la qualité des résultats obtenus est comparable à celle observée chez d'autres groupes.

Ensuite, nous avons appliqué notre procédure à des systèmes plus complexes tels que les systèmes protéine:ligand. Nous nous sommes intéressés à l'aspartyl-ARNt synthétase (AspRS) et l'asparaginyl-ARNt synthétase (AsnRS). Ces enzymes jouent un rôle crucial dans la traduction du code génétique. Les synthétases fixent leur acide aminé spécifique sur leur ARNt correspondant établissant ainsi l'intégrité du code génétique. Tout d'abord nous avons réalisé le 'design' des sites actifs d'AspRS et d'AsnRS en présence de leur ligand natif et non natif afin d'évaluer les performances de notre procédure. La qualité des séquences prédites est comparable à celle observée pour les protéines globulaires entières. Par ailleurs, nous avons montré que notre procédure était sensible à la nature du ligand présent dans la poche. Enfin, nous avons réalisé le 'design' d'un nombre limité de positions dans le site actif de l'AsnRS de façon à ce qu'elle lie préférentiellement l'aspartate au détriment de l'asparagine. Un jeu de mutants prometteurs fut retenu. Leur stabilité et affinité pour les ligands natifs et non natifs est actuellement analysé par des simulations de dynamique moléculaire.

ABSTRACT

Protein design aims to develop new proteins with new structural and/or functional properties. The principle is to identify among the available sequences, those that preserve the protein's three dimensional fold and that confer the desired properties. The general procedure can be decomposed into two steps : (i) the interaction energy between all the residue pairs is precomputed and stored in an energy matrix taking into account all amino acid types and all possible conformations, (ii) an optimization algorithm explores simultaneously sequence and conformational space to determine the best amino acid combination. Next, different filters (based on affinity, protein stability…) can be applied to separate functional sequences (for a given fold) from the non functional ones.

We first focused on the development of the protein design procedure, particularly, on the setting up and optimization of the energy function and the implementation of the optimization algorithm. We have shown that our procedure is robust because it performs well for a wide variety of applications crucial in protein design such as: prediction of sidechain orientation, prediction of stability or affinity changes due to point mutations and design of native-like sequences for a set of globular proteins. For all of these applications the quality of results is competitive with those obtained by other groups.

Next, we applied our procedure to more complex systems such as protein:ligand complexes. We focused on aspartyl-tRNA synthetase (AspRS) and asparaginyl-tRNA synthetase (AsnRS). These enzymes play a crucial role in preserving the accuracy of genetic code translation, linking their specific amino acid to a cognate tRNA, which carries the corresponding anticodon. First, we performed the design of the whole active site of AspRS and AsnRS in presence of their specific or non specific ligands in order to test the performance of our procedure. The quality of the designed sequences is consistent with those observed on entire globular proteins. On the other hand, we showed that our procedure was sensitive to the nature of the ligand present in the active site. Finally, we performed the design of a limited number of selected positions of the AsnRS active site so that the synthetase can bind preferentially aspartate over asparagine. A set of promising mutants has been retained. Their stability and affinity for native and non-native ligands are now analyzed by molecular dynamics.

Table des matières

Chapitre 1

Protein design

La conception des protéines ou 'protein design' consiste à développer de nouvelles protéines dont on aurait modifié les propriétés structurales et/ou fonctionnelles. Cette discipline récente peut être abordée par des approches expérimentales ou computationnelles. Deux stratégies prédominent dans l'approche expérimentale. La première, dite 'rationnelle', repose sur la connaissance de la structure 3D de la protéine d'intérêt et consiste à muter des régions particulières de la protéine en vue d'obtenir les modifications désirées. Cette approche s'appuie sur les techniques aujourd'hui bien maîtrisées de mutagenèse dirigée et présente ainsi l'avantage d'être relativement facile à mettre en place. La seconde méthode, dite 'méthode combinatoire', consiste à muter aléatoirement la protéine cible et à mettre en place une procédure de sélection de façon à identifier les mutants présentant les propriétés désirées. Cette méthode mime les mécanismes de l'évolution naturelle et est connue sous le nom d'évolution dirigée. A l'inverse de la première approche, l'évolution dirigée ne requiert ni la connaissance de la structure 3D de la protéine ni de comprendre le mécanisme de la protéine. D'ailleurs, il arrive avec cette méthode d'obtenir la propriété désirée en introduisant des mutations complètement inattendues. Cette technique est couramment utilisée depuis quelques années pour développer

de nouveaux ligands, biosenseurs et catalyseurs et joue un rôle de plus en plus important dans les domaines pharmaceutiques et bio-technologiques. Néanmoins, cette approche s'avère coûteuse, longue à réaliser et ne permet pas une exploration complète de l'espace des séquences disponibles pour la région à muter.

En revanche, l'approche computationnelle, permet une exploration plus exhaustive et contrôlée des mutations possibles plutôt que de générer expérimentalement et aléatoirement des librairies de mutants. Cette approche est connue en anglais sous le nom de 'Computational Protein Design'. Nous y ferons référence par l'abbréviation CPD. Par ailleurs, la notion de 'design' n'a pas de correspondance exacte en français, aussi nous conserverons le terme anglais tout au long de ce manuscrit. Le CPD fut au départ essentiellement appliqué au renforcement de la stabilité de protéines existantes, le développement de protéines thermostables présentant un intérêt industriel indéniable. Par la suite, les applications du CPD furent étendues au problème inverse du repliement, au développement de nouveaux chemins de repliements ou encore de protéines présentant de nouvelles fonctions[116]. L'ensemble de ces différentes applications sera présenté en section 1.5.

L'un des objectifs les plus courants du CPD consiste à identifier, parmi toutes les séquences possibles, celles capables de se replier dans une structure donnée. Le nombre de séquence possible est énorme. Pour simplifier le problème, on commence par discrétiser l'espace conformationnel, en utilisant la notion de rotamères, par exemple. La procédure générale est ensuite réalisée en deux étapes. La première consiste à calculer une matrice d'énergie contenant les énergies d'interactions entre toutes les paires de résidus de la protéine en autorisant successivement tous les types d'acides aminés dans toutes leurs conformations possibles. Cette matrice d'énergie contiendra aussi l'énergie d'interaction de chaque résidu avec le squelette peptidique. La seconde étape, ou 'phase d'optimisation', consiste à déterminer la combinaison optimale d'acide aminés étant donné le repliement protéique d'intérêt.

Ainsi, le CPD nécessite plusieurs ingrédients. Il faut une fonction d'énergie et un algorithme d'optimisation capables de distinguer parmi l'espace des séquences disponibles, celles qui confèreront les propriétés structurales et/ou fonctionnelles attendues à la protéine d'intérêt. Cette fonction d'énergie devra inclure un modèle de solvant fiable et peu coûteux. Par ailleurs, le CPD requiert une description correcte de l'état replié mais aussi de l'état déplié afin de déterminer les séquences les plus favorables. Enfin cette procédure devra être validée par un ensemble de tests.

1.1 Représentation du système

1.1.1 Représentation de l'état replié par une discretisation de l'espace conformationel

Comme nous venons de le voir, l'un des objectifs les plus courants du CPD consiste à identifier parmi l'espace des séquences possibles, celles qui préserveront le repliement de la protéine d'intérêt. Or pour une petite protéine de 100 résidus en permettant les 20 acides aminés classiques à chaque position nous obtenons déjà un nombre potentiel de 20^{100} séquences. Ce nombre peut encore être augmenté si on tient compte des différentes orientations possibles des chaînes latérales et des différentes conformations de la chaîne principale. C'est pourquoi dans le but de réduire la complexité de l'espace conformationel le CPD impose une discrétisation de ce dernier. Cette discrétisation a lieu a deux niveaux :

- Discrétisation de l'espace conformationel des chaînes latérales
- Discrétisation de l'espace conformationel de la chaîne principale

Discrétisation de l'espace conformationel des chaînes latérales

En effet chaque acide aminé peut être présent sous différentes conformations. La géométrie des chaines latérales peut être définie par des angles de torsion nommés

par convention χ_1, χ_2,..., en partant de la chaine principale jusqu'à l'extrémité de la chaine latérale. Ces angles correspondent à la rotation des groupes chimiques autour des liaisons (figure 1.1).

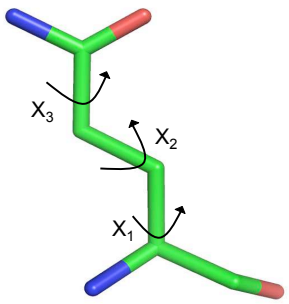


FIG. 1.1 – **Angles dièdres χ_1, χ_2 et χ_3 de la glutamine.**

L'espace conformationnel peut alors être réduit en un nombre fini d'angles de torsion, adoptant une série finie de valeurs. Cette discrétisation de l'espace se prète bien aux protéines puisqu'en pratique, pour les acides aminés, certaines valeurs d'angle de torsion sont nettement plus probables que d'autres. En effet dans des travaux pionniers, Ponder and Richards ont montré que dans les structures cristallines protéiques les chaînes latérales adoptaient un nombre limité de conformations préférentielles [168]. Ces conformations sont aussi connues sous le nom de 'rotamères', concept introduit par Janin et coll. en 1978 [97] (figure 1.2). Chaque type d'acide aminé présente typiquement deux ou trois angles de torsions ce qui conduit à une moyenne d'environ 10 rotamères préférentiels. Il existe cependant une faible proportion estimée inferieure à 5% de conformations mal représentées par les rotamères. L'approximation rotamèrique introduit donc une légère erreur dans la modélisation des protéines ; néanmoins, cette description discrète permet de réduire considérablement l'espace conformationel, avantage essentiel pour le CPD.

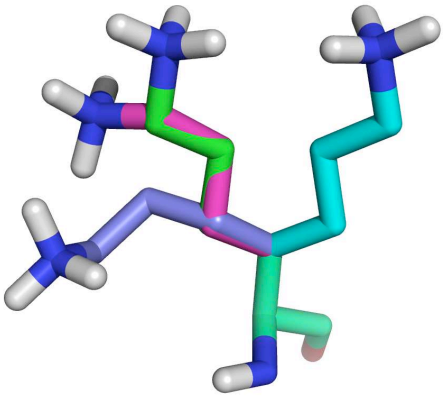


FIG. 1.2 – **Représentation de quatre rotamères de la lysine.** Les cordonnées des atomes N, H, Cα, Cβ, C et O sont communes à tous les rotamères, tandis que les coordonnées de la chaîne latérale varient d'un rotamère à l'autre.

Discrétisation de l'espace conformationel de la chaîne principale

Dans de nombreuses études, la chaîne principale est maintenue fixée dans sa conformation native afin de simplifier l'exploration de l'espace conformationnel [43], [54], [118], [211]. La fixation du squelette peptidique a fait ses preuves dans certaines applications telles que la génération de protéines hyperstables [139] ou le *de novo design* d'une protéine complète [44]. Cependant, cette approximation rencontre des limites. Par exemple, certaines chaînes latérales peuvent être considérées comme défavorables énergetiquement alors qu'un léger ajustement du squelette aurait suffit pour diminuer leur energie considérablement.

Introduire de la flexibilité dans le squelette augmente considérablement l'espace conformationel. Deux approches prédominent. La première consiste à générer un ensemble de squelettes et à optimiser la séquence d'acide aminés à partir de cet

ensemble de squelettes maintenus fixes durant l'optimisation [195], [124], [121]. La seconde consiste à réajuster la chaîne principale pour un grand nombre de séquences fixes [84]. Les deux approches nécessitent de spécifier à l'avance un nombre limité de conformations de la chaîne principale. A partir de ces deux approches, différentes études furent réalisées avec plus ou moins de succès.

La première fut effectuée par le groupe de Harbury en 1998 qui s'intéressa aux protéines présentant une symétrie dans leur structure telles que les *coiled-coils* ou *TIM barrels*. Dans ces cas, leur squelette protéique peut être décrit par des équations paramétriques [36], [157], [158]. La symétrie réduit considérablement le nombre de conformations de la chaîne principale. C'est ainsi que Harbury et coll. modélisèrent une famille de dimères, trimères et tetramères de *right-hand coiled-coils*, repliements jamais observés dans la nature [84]. Cependant cette approche ne peut être généralisée à la grande majorité des repliements qui en général, ne présentent pas de symmétrie.

Su et Mayo abordèrent le problème différemment en traitant les éléments de structure secondaires comme des corps rigides capables de se déplacer les uns par rapport aux autres [195]. Ils réussirent à redessiner des variants stables de la protéine Gβ1. Ils montrèrent alors que pour de légeres perturbations de la chaîne principale, les séquences prédites étaient identiques à celles obtenues avec un squelette rigide. Ceci suggère que les fonctions d'energie courantes ne sont pas sensibles à de subtils changements du squelette.

Desjarlais et Handel modélisèrent explicitement la flexibilité du squelette en utilisant la combinaison d'un algorithme génétique et d'un échantillonage Monte Carlo. De nombreuses conformations de la chaîne principale purent ainsi être échantillonées aboutissant au *design* de trois variants stables du coeur protéique de 434 cro [55]. Toutefois, les auteurs notèrent que les résultats de prédiction de changements de

stabilité associés à des mutations ponctuelles de la protéine étaient indépendants de l'utilisation d'un squelette rigide ou flexible. Cette observation confirme le manque de sensibilité des fonctions d'énergies actuelles à de subtils changements conformationnels du squelette.

Kono et Saven utilisèrent des ensembles de structures RMN afin de représenter différentes conformations de la chaîne principale [113]. Le groupe de Baker connu aussi de nombreux succès dans ce domaine. Leur méthode consiste à échantillonner alternativement l'espace des séquences et des structures en optimisant itérativement les séquences pour un squelette peptidique fixe puis la conformation du squelette étant donnée une séquence d'acides aminés [120], [180]. Cette approche permit en particulier à Kuhlman et coll. de créer un nouveau repliement protéique jusqu'alors absent de la Protéine Data Bank [11], [120]. Saunders et Baker montrèrent ensuite que cette méthode pouvait être utilisée pour mimer certains aspects de l'évolution. Ainsi, ils furent capables d'échantillonner la diversité de séquences de certaines familles protéiques avec une plus grande exactitude que lors de l'utilisation d'un squelette fixe ou d'ensemble de squelettes présentant des perturbations aléatoires [180].

Récemment, Fu et coll. démontrèrent que l'utilisation des modes normaux permettait également de simuler la flexibilité du squelette peptidique. En particulier, ces derniers appliquèrent cette méthode au *design* de peptides interagissant avec le récepteur Bcl-x_L (membre de la famille Bcl-2). Partant de la structure cristallographique du complexe Bcl-x_L/Bim, les auteurs générèrent par des calculs de mode normaux, un ensemble de conformations du ligand peptidique et montrèrent que cette approche conduisait à un jeu de solutions de basses énergies plus grand et plus divers qu'en se limitant à la structure cristallographique du complexe [71].

1.1.2 Représentation de l'état déplié

Lors de son processus de repliement, une protéine va échantillonner un ensemble de conformations différentes toutes moins structurées et compactes que la native jusqu'à se replier dans sa conformation la plus stable, appelée conformation native. Sa stabilité est mesurée par la différence d'énergie libre entre sa structure native et l'ensemble des états non-natifs. Bien que la structure d'une protéine ne soit pas figée, cette dernière adoptera sa conformation native la plus grande partie du temps. D'un point de vue thermodynamique, ce temps augmente de façon exponentielle avec l'énergie libre de repliement de la protéine. Pour un squelette peptidique donné, la séquence la plus favorable correspond à celle qui maximisera la différence d'énergie libre entre l'état replié et l'état déplié.

Modéliser l'état déplié est loin d'être évident puisque sa caractéristique est de ne pas être très structuré. L'état déplié consiste en une distribution continue de conformations ou micro-états d'énergies similaires. Ces différentes conformations conduisent à l'exposition au solvant au moins partielle des résidus hydrophobes de la protéine. Aujourd'hui, l'état déplié demeure très difficile à caractériser par les méthodes expérimentales classiques. Certaines approches telles que le dichroïsme circulaire ou la spectroscopie infra-rouge peuvent nous renseigner sur la présence d'éventuelles structures secondaires résiduelles. Toutefois, aucune méthode actuelle n'est en mesure d'apporter une information structurale pour l'état déplié avec un même niveau de détail que pour l'état natif.

Ainsi, différents modèles furent proposés pour représenter l'état déplié. Un des modèles les plus simples décrit l'état déplié par une chaîne polypeptidique étendue si bien que les chaînes latérales des acides aminés interagissent essentiellement avec le solvant et les groupes voisins du squelette peptidique. En revanche, les chaînes latérales n'engagent que très peu d'interactions entre elles. Par conséquent, ce modèle considère que l'énergie libre de l'état déplié est uniquement dépendante de la composition en acides aminés et non de la séquence. L'approche la plus courante consiste à déterminer des énergies de référence pour chaque type d'acide aminé, représentant ainsi sa contribution individuelle à l'énergie libre de l'état déplié. Cette situation peut être modélisée par une collection de n tripeptides de séquence ala-X-ala où n correspond au nombre de résidus dans la protéine et X à l'acide aminé courant d'une position donnée [43], [211].

Par ailleurs, le groupe de Shortle émet l'hypothèse que les protéines dépliées engageraient des interactions intra-chaînes conduisant à des structures locales absentes du modèle de tripeptide [162]. Ainsi, Handel et coll. utilisèrent des fragments de 13 résidus extraits de structures natives présentant différents éléments de structure secondaire [165]. Cependant, cette approche plus sophistiquée et plus lourde n'apporte pas d'améliorations significatives par rapport au modèle des tripeptides. Les auteurs ont en effet, testé les capacités de cette méthode à prédire des changements de stabilité associés à des mutations ponctuelles. L'erreur moyenne obtenue avec leur méthode est de 1.05 kcal/mol contre 1.20 kcal/mol avec le modèle des tripeptides [165]. Le modèle des tripeptides demeure aujourd'hui, le modèle le plus communément utilisé.

1.2 Fonctions d'Energie

1.2.1 Potentiels de mécanique moléculaire classique à l'origine de ceux du CPD

Aujourd'hui, la plupart des modèles utilisés pour l'étude des protéines s'appuient sur les principes de la mécanique moléculaire. Cette description moléculaire utilise des concepts physiques simples hérités de Newton, Coulomb et Laplace. Les atomes de la protéines sont représentés par des particules sphériques avec un rayon plus ou moins incompressible et une charge nette constante dérivée de calculs de mécanique quantique ou de résultats expérimentaux. Les liaisons inter-atomiques sont considérées comme des petits ressorts avec une longueur d'équilibre équivalente aux longueurs de liaisons déterminées expérimentalement [146], [18].

La dynamique moléculaire consiste à calculer l'évolution d'un système de particules au cours du temps. Elle requiert une fonction d'énergie capable de décrire les forces qui guideront les différents atomes de la protéine durant la simulation et s'appuie principalement sur des potentiels physiques connus sous le nom de 'champ de force'. Ces potentiels sont principalement derivés de calculs de mécanique quantique et thermodynamiques, ainsi que de données cristallographiques et spectroscopiques obtenues à partir d'un grand nombre de systèmes différents. Plusieures années de recherche furent nécessaires pour paramétriser ces potentiels déstinés à la simulation de protéines. Ensuite, d'autres potentiels tels que les potentiels statistiques furent aussi incorporés dans les fonctions d'énergie. Ces derniers sont déduits de bases de données de structures protéiques connues. Un des avantages de ce type de fonction d'énergie est qu'il peut modéliser n'importe quel comportement déjà observé et intégré dans ces bases de données même si les mécanismes physico-chimiques de ce comportement demeurent incompris. Cependant le grand désavantage de ces fonctions d'énergie est qu'elles sont incapables de prédire de nouveaux comportement absents des bases de données.

La première simulation de dynamique moléculaire d'une protéine fut réalisée par Martin Karplus et ses collaborateurs en 1977 [145]. Ils simulèrent une petite protéine sur un temps de quelques picosecondes étant donné les puissances de calculs disponibles à l'époque. Depuis, la dynamique moléculaire est communément exploitée pour simuler des gros systèmes protéiques sur des temps bien plus conséquents, affiner des structures cristallines ou encore reproduire des processus de repliement protéique.

Toutefois, si ces types de modèles sont appropriés pour les simulations de dynamique moléculaire, le CPD impose de considérables modifications de la fonction d'énergie. En effet, cette dernière doit faire un compromis entre exactitude et rapidité. Par conséquent, les algorithmes de CPD utilisent des fonctions d'énergies généralement moins lourdes d'un point de vue computationel que celles exploitées en dynamique moléculaire [13], [81], [150], [166]. En particulier, la discrétisation de l'espace conformationnel exige des ajustements empiriques de la fonction d'énergie que nous détaillerons par la suite.

Par ailleurs, certains algorithmes d'optimisation imposent des termes énergétiques décomposables en somme d'interaction de paires. C'est ainsi que dans leur forme la plus standard, les fonctions d'énergie en CPD sont constituées de termes énergétiques dits 'de paires' tels que van der Waals, électrostatique, liaisons hydrogènes et termes de surface (la décomposition des termes de surfaces en somme de paires est présentée dans les sections 1.2.2 et 4.1.4). Ainsi, l'énergie globale de la protéine sera décomposée en une somme d'énergies d'interactions de paires de résidus et d'interactions entre les différents résidus et le squelette peptidique. Ces énergies seront alors stockées dans la matrice d'énergie mentionnée précédemment.

Enfin, les différentes contributions de la fonction d'énergie sont typiquement calibrés et pondérés empiriquement pour optimiser la performance d'un type d'application en particulier. Ainsi, alors que les différents champs de force en dynamique moléculaire restent très standards, les potentiels en CPD varient remarquablement d'une équipe à l'autre.

1.2.2 Electrostatique et solvatation

Un des buts initiaux du CPD était de renforcer la stabilité de protéines d'intérêt [41]. Aujourd'hui, le CPD se tourne vers le *design* de nouvelles fonctions. Les interactions électrostatiques prennent alors une grande importance [205]. Les résidus polaires ou chargés sont généralement présents à la surface des protéines et engagent un grand nombre d'interactions électrostatiques favorables avec les molécules d'eau du solvant. Toutefois, on relève la présence de quelques résidus polaires ou chargés dans le coeur des protéines. Ces derniers sont connus pour être déstabilisants du fait de leur désolvatation fortement défavorable [104]. On peut alors supposer que ces résidus jouent un rôle important pour la fonction et/ou le repliement final de la protéine. Ainsi, ils limitent le nombre total de conformations possibles à celles permettant des interactions suffisamment favorables (pont salins, liaisons hydrogènes). *A contrario*, une protéine dont le coeur serait uniquement composé de résidus hydrophobes ne présenterait pas nécessairement un repliement unique. C'est l'observation que fit le groupe d'Eisenberg [89]. Ils conçurent un peptide synthétique de 16 résidus se repliant sous la forme d'une hélice α amphipathique. Ils montrèrent que cette hélice pouvait être présente aussi bien sous forme de dimères, tétramères ou héxamères, chacun des modes d'association impliquant un grand nombre d'interactions hydrophobes.

Lumb et Kim, mirent également en évidence le rôle déterminant des résidus polaires enfouis dans l'obtention d'une structure unique [136]. Ils s'intéressèrent à deux peptides synthétiques ACID-p1 et BASE-p, chacun se repliant en une hélice α amphipathique. Les deux peptides forment un hétérodimère *coiled-coil* parallèle et leur interface est exclusivement composée de résidus hydrophobes (leucines) à l'exception d'un unique acide aminé polaire (asparagine) présent dans chacun des

deux peptides. L'asparagine du peptide ACID-p1 réalise alors une liaison hydrogène avec celle du peptide BASE-p1. Les auteurs étudièrent alors le rôle de cette interaction polaire en mutant ces deux asparagines par deux leucines. Les mutants présentaient alors une stabilité nettement supérieure à celle des deux peptides p1 mais formaient préférentiellement des hétérotétramères au lieu d'un hétérodimère. Par ailleurs, l'hétérotétramère ne formait pas de repliement unique ; en particulier les hélices présentaient différentes orientations parallèles ou anti-parallèles.

Par ailleurs, Shoichet et coll. s'intéressèrent au compromis existant entre la stabilité d'une protéine et la réalisation de sa fonction [189]. Ils réalisèrent des mutations ponctuelles de résidus du site actif du lysozyme T4 et montrèrent qu'elles résultaient en une augmentation de la stabilité thermique des mutants couplée systématiquement à une perte d'activité de ces derniers. Shoichet et coll. concluent alors que les résidus contribuant à la fonction de la protéine (activité catalytique, interaction protéine-ligand...) par l'intermédiaire de ponts salins ou liaisons hydrogènes ne sont pas optimisés pour maximiser la stabilité de celle-ci.

Toutes ces études soulignent le rôle des interactions électrostatiques dans l'unicité du repliement et dans la réalisation de la fonction des protéines. Par conséquent le *design* de nouvelles fonctions protéiques ou de nouveaux repliements nécessite des modèles physiques capables de reproduire correctement les forces électrostatiques permettant à une protéine de se replier et d'être fonctionnelle. Une protéine est le plus souvent au contact d'un milieu acqueux qui est très polarisable. Les molécules d'eau peuvent ainsi intéragir directement avec des résidus hydrophiles de la protéine, entrant en compétition avec les autres groupes polaires environnants. Le solvant a alors un effet 'écran' sur les interactions électrostatiques. Une description correcte des interactions électrostatiques implique donc la prise en compte du solvant dans la fonction d'énergie. Le modèle doit être à la fois suffisamment fiable et peu gourmand en temps de calcul.

En utilisant un modèle électrostatique simple, Marshall et coll. réussirent à redessiner une protéine plus stable en en imposant des restrictions sur la charge des résidus dans les régions N-terminale et C-terminale des hélices de la protéine [141],[216] . Une autre étude réalisée sur une base tout aussi empirique permit la conception d'un variant de thioredoxine plus stable [14]. Lors du calcul de CPD, tous les rotamères d'acides aminés polaires n'engageant pas un nombre minimal de liaisons hydrogènes furent éliminés. Bien que ces approches empiriques aient faire leur preuves dans quelques cas isolés il apparut comme crucial de développer des champs de force plus généraux, basés sur des modèles physiques, capables de reproduire la balance entre le phénomène de solvatation et celui des interactions électrostatiques.

Ainsi, différents modèles furent introduits dans les fonctions d'énergie de CPD pour décrire la contribution électrostatique et celle du solvant. Tous ces modèles ont la caractéristique de représenter les molécules d'eau de manière implicite, l'utilisation de modèles de solvant explicite étant beaucoup trop coûteuse. Il existe deux principaux types de modèles implicites : les modèles 'empiriques' et les modèles 'continus'. Les modèles empiriques sont dépendants de paramètres de solvatation directement déduits de données expérimentales. L'énergie libre de solvatation du soluté est considérée comme la somme des contributions de tous ses atomes. La contribution de chaque groupe est généralement approximée par une fonction linéaire de sa surface accessible au solvant [61],[212], [164], [70], [75], [74]. Ces termes incluent les composantes hydrophobes et électrostatiques de la solvatation mais pas l'effet 'écran' du solvant sur les interactions électrostatiques entre différents résidus chargés. Cet effet doit alors être introduit comme un terme additionnel. Les modèles continus définissent le soluté et le solvant comme deux régions distinctes caractérisées par deux constantes diélectriques différentes. L'énergie libre de solvatation est obtenue en résolvant l'équation de Poisson-Boltzmann [107],[93]. Toutefois, comme nous le verrons par la suite la résolution de cette équation par des méthodes numériques

s'avère coûteuse et d'autres modèles traitant le problème de façon analytique durent être implémentés pour pallier cette difficulté.

Modèles empiriques

Modèle CASA (Coulombic Accessible Surface Area)

Sa simplicité et son efficacité font de lui le modèle le plus répandu dans le domaine du CPD. Initialement implémenté par Wesson et Einsenberg [212], ce modèle utilise une constante diélectrique afin de mimer l'écrantage des interactions électrostatiques protéine-protéine due à la présence du solvant très fortement polarisé. A ce terme électrostatique, il faut ajouter un terme dépendant de la surface accessible au solvant. Les différents types atomiques sont alors caracterisés par des paramètres de solvatation atomique dérivés de calculs expérimentaux d'énergie libre de transfert octanol-eau ou vapeur-eau. Ces paramètres vont refléter le caractère hydrophile/hydrophobe de chaque type atomique. La contribution énergétique de chaque atome est donnée par le produit de son paramètre de solvatation atomique et de sa surface accessible au solvant. Ce terme va favoriser l'exposition des groupes polaires et pénaliser leur enfouissement.

Pour être utilisé dans le cadre du CPD, ce modèle doit satisfaire le prérequis d'une fonction d'énergie dite de 'paires'. En particulier, l'énergie de solvatation de la protéine doit être décomposable en termes d'énergie de paires de résidus. Ceci pose un problème dans le cas du calcul des surfaces enfouies des résidus de la protéine puisque ces surfaces dépendent de la configuration globale de la protéine. En effet, avec les algorithmes habituels, un point à la surface d'un atome peut être au contact de deux autres atomes simultanément. Une approche de 'paires' aboutirait alors à une sur-estimation de la surface enfouie de l'atome (figure 1.3). Street et Mayo développèrent une approximation permettant un calcul de 'paires' des surfaces enfouies des différents résidus de la protéine [193].

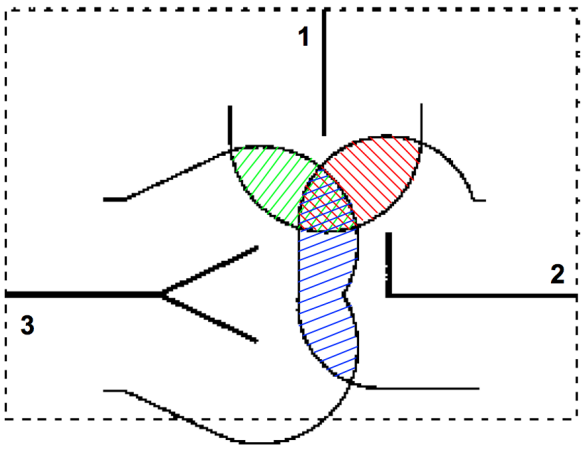


FIG. 1.3 – **Représentation de trois résidus et de leurs surfaces de contact respectives.** En rouge : la surface de contact entre les résidus 1 et 2. En vert : la surface de contact entre les résidus 1 et 3. En bleu : la surface de contact entre les résidus 2 et 3. L'intersection correspond à la surface de contact commune aux trois résidus, ainsi cette surface est sur-estimée puisqu'elle sera comptée deux fois avec les algorithmes d'optimisation classiques.

Malgré sa simplicité, le modèle CASA a largement démontré son utilité dans le développement de protéines plus stables ou encore dans le *design* de coeurs hydrophobes [43],[170]. En revanche, il est moins adapté au *design* de surfaces protéiques. Des modèles electrostatiques plus sophistiqués furent développés par la suite.

Modèle Lazaridis & Karplus (LK)

Proposé par Lazaridis et Karplus, ce modèle considère que l'énergie de solvatation d'un atome dépend du nombre de molécules d'eau exclues par les résidus environnants lors du processus de repliement [128]. Cette énergie est calculée comme le produit du volume de désolvatation de l'atome en question par la densité d'énergie libre de solvatation autour de cet atome. Cette densité d'énergie décroît rapidement en s'éloignant du soluté. Ainsi, les atomes présents dans la première couche de solvatation ont un effet de désolvatation beaucoup plus grand que les atomes plus éloignés. Finalement, l'énergie de solvatation d'une protéine correspond à une somme de contributions de groupes atomiques. Ces dernières sont déterminées à partir de mesures expérimentales d'énergie libre de solvatation réalisées sur des petites molécules. Ce modèle présente donc l'avantage de ne pas recourir à des calculs de surface accessible au solvant et permet ainsi un gain de temps notable. Kuhlman et

Baker l'appliquèrent avec succès au *design* d'un nouveau repliement protéique [120].

Modèles continus

Le modèle de Poisson-Boltzmann (PB)

Actuellement considéré comme le meilleur modèle de solvant implicite, le modèle de Poisson-Boltzmann est entièrement fondé sur des concepts physiques. Il considère le soluté (la protéine) comme une cavité de faible constante diélectrique incorporée dans un milieu de forte constante diélectrique. Ce modèle repose essentiellement sur deux ingrédients physiques : (1) les fortes interactions électrostatiques entre groupes chargés et solvant polarisé, (2) le phénomène d'écrantage du solvant sur les interactions intra-protéine. La polarisation induite dans le solvant est alors utilisée pour déterminer le potentiel électrostatique au sein de la protéine. Initialement, des solutions analytiques étaient suffisantes pour son application à des systèmes très simples tels que des solutés sphériques ou cylindriques. Cependant, des méthodes numériques plus lourdes durent être implémentées afin de résoudre l'équation de Poisson-Boltzmann pour des systèmes plus complexes tels que les protéines.

Si ce modèle s'appuie principalement sur les concepts physiques macroscopiques des milieux diélectriques, il a démontré sa capacité à reproduire des énergies de solvatation à l'échelle microscopique des protéines. En effet, il s'est montré très efficace[1] dans des problèmes de liaison de ligand [5], d'interactions protéine-protéine [88],[79], de prédiction de pKa [8], de dynamique moléculaire [48],[69], [135] ou encore dans la simulation de protéines membranaires [178]. En revanche, le modèle de Poisson-Boltzmann reste difficilement applicable au problème du CPD puisqu'il n'est pas décomposable en énergie de paires.

[1] avec des résultats dignes des simulations réalisées en solvant explicite

En effet, avec PB, l'interaction d'une paire d'acides aminés dépend de la forme des régions de haute et basse constantes diélectriques et donc des coordonnées de la protéine entière. Considérer la protéine dans sa globalité n'est cependant pas réalisable pour chacune des combinaisons de séquences possibles testées dans les calculs de CPD. Marshall et coll. [142] ont alors proposé une nouvelle formulation du modèle PB décomposable en énergie de paires de chaînes latérales. Ils utilisent une représentation réduite de la protéine fondée sur le squelette peptidique et une ou deux chaînes latérales afin d'approximer l'environnement diélectrique au sein et autour de la protéine. L'énergie électrostatique pour chaque chaîne latérale ou paire de chaînes latérales peut alors être déterminée en réalisant des calculs de PB/FDPB à partir de cette représentation simplifiée. Cette approximation est capable de reproduire des effets résultants de changements d'acides aminés en surface. En outre la méthode produit des énergies tout à fait comparables à celles obtenues avec la représentation complète de la protéine. Ceci est encourageant pour le problème du CPD si l'on limite son application à de petits systèmes, le modèle de PB demeurant coûteux en temps de calcul. On peut toutefois lui attribuer un rôle très utile pour le calibrage et la validation des autres modèles de solvant.

Le modèle de Tanford-Kirkwood modifié

Havranek et Harbury reprirent le modèle de Tanford-Kirkwood en vue d'augmenter sa rapidité d'exécution et ainsi pouvoir étendre son application au domaine du CPD [85], [197]. La protéine est assimilée à une sphère de faible constante diélectrique permettant la résolution de l'équation de Poisson-Boltzmann par des calculs analytiques. Les charges individuelles sont cartographiées à partir de leur coordonnées dans la protéine sur la sphère. Cette sphère est incorporée dans un milieu de forte constante diélectrique. Les énergies d'interactions entre les différentes charges sont enfin calculées analytiquement selon l'équation de Poisson-Boltzmann. Cette méthode fut appliquée avec succès pour le *design* de repliements *coiled-coil* capables de se dimériser [84].

Le modèle de Born Généralisé (GB)

Le modèle de Born Généralisé reprend le même concept que celui de Poisson-Boltzmann en modélisant la protéine comme une cavité entourée d'un continuum diélectrique jouant le rôle de solvant. En utilisant quelques approximations le modèle GB permet la résolution de l'équation de *Poisson* par une approche analytique rendant par conséquent son temps d'exécution beaucoup plus rapide. Sa particularité, est de représenter chacune des charges de la protéine au centre d'une sphère dont le rayon, dit 'rayon de Born', représente la distance entre la charge et le solvant, reflétant ainsi l'enfouissement de cette dernière dans la protéine. Le potentiel électrostatique de chaque charge peut alors être déterminé. Une première contribution consiste en l'énergie propre de chacune des charges atomiques. Cette énergie est le reflet de l'interaction de la charge avec le solvant. La seconde contribution correspond à l'interaction de cette même charge avec toutes les autres charges environnantes en tenant compte du phénomène d'écrantage des interactions électrostatiques par le solvant.

L'efficacité du modèle de GB dépend essentiellement de l'exactitude avec laquelle ont été calculés les rayons de Born. On compte aujourd'hui de nombreuses variations du modèle de GB différant par la façon dont sont déterminés ces rayons. Son aptitude à reproduire des résultats comparables à ceux du modèle de PB dépend fortement des différentes implémentations de celui ci.

Le modèle de GB a fait ses preuves dans des applications très variées telles que des calculs de solvatation [163], [57], prédiction de pKa [190],[131], simulation de protéines et d'acides nucléiques par dynamique moléculaire [24], [57], prédiction de boucles [172], problème d'interaction protéine-ligand [138], [5], classement de repliements natifs parmi des non natifs [58].

Toutefois, commme pour le modèle de PB, le Born Généralisé requiert quelques modifications pour devenir décomposable en énergie de paires et ainsi pouvoir être appliqué au CPD. Récemment, Pokala et Handel mirent en place une approximation afin de reproduire l'environnement local durant l'étape de calcul d'énergie d'une paire de résidus [167]. Ils utilisèrent des pseudo-résidus sphériques pour représenter les chaînes latérales environnantes (d'identité inconnue à ce stade du calcul). Cette méthode se révèle beaucoup plus rapide que Poisson-Boltzmann tout en conservant sa fiabilité. En particulier, Pokala et Handel l'utilisèrent avec succès pour la prédiction de pKa sur un jeu de plus de 200 groupes ionisables provenant d'une quinzaine de protéines différentes [167]. Archontis et Simonson approximèrent le modèle classique de GB par une approche originale [4]. Les résidus sont considérés comme des unités à part entière puisqu'on ne définit plus des rayons de solvatation atomiques mais de résidus (Bi). L'énergie 'propre' de chaque résidu est décomposable en terme d'énergie de paires de résidus ce qui n'est pas le cas du terme énergétique qui décrit l'effet écran du solvant. Ce terme dépend de la configuration globale de la protéine. Néanmoins, les auteurs montrent qu'il peut être approximé par une fonction parabolique de B, B étant le produit des Bi, Bj d'une paire de résidus i, j. Cette expression dépend uniquement de la paire de résidus i, j et non de son environnement et reproduit correctement l'environnement diélectrique de la paire en question. Cette méthode apporte des résultats comparables au modèle de GB classique et se révèle très encourageante pour les problèmes de CPD.

Nous avons présenté les principaux modèles de solvant implicite. Le Poisson-Boltzmann est actuellement le modèle de référence pour sa capacité à reproduire des résultats dignes de ceux obtenus avec des solvants explicites. Les modèles empiriques tels que celui de CASA sont actuellement les plus simples et les plus rapides mais peinent généralement dans le *design* des résidus de surface [149]. Différentes études se sont attelées à la comparaison des différents modèles de solvant implicites. En

particulier, Jaramillo et Wodak ont comparé leur aptitude à favoriser ou non l'enfouissement de résidus polaires, à reproduire des séquences proches de la native lors de calculs de CPD ou encore à discriminer des repliements natifs de non-natifs [99]. Les auteurs concluent que seul le modèle empirique de solvatation atomique est capable de limiter l'enfouissement de résidus polaires et de reproduire des séquences de type natif, les autres modèles se révélant à l'heure actuelle inadéquates. Toutefois, leurs résultats restent difficilement interprétables. En effet, l'analyse de leur protocole révèle une mauvaise paramétrisation des modèles de GB utilisés. De plus, leur protocole résulte en la formation de cavités artificielles dans la protéine qui faussent les résultats des calculs Poisson-Boltzmann. Par ailleurs, il faut noter qu'il existe un grand nombre de variants du GB et que seules deux versions furent testées dans cette étude. En revanche cette étude met en évidence la complexité des modèles de GB, l'importance du choix de paramètres et les difficultés rencontrées pour leur bonne utilisation. Cependant, elle ne permet en aucun cas la condamnation des modèles continus tels que PB ou GB. Dans tous les cas, on peut conclure que le CPD représente un bon défi pour l'évaluation des différents modèles de solvant implicites.

1.2.3 Energie de van der Waals

La force de van der Waals correspond à une interaction de faible intensité entre les atomes d'une protéine. Elle est composée de deux termes : un terme attractif dominant à grande distance et un terme répulsif dominant à courte distance. Le potentiel Lennard-Jones est une approximation mathématique utilisée couramment en modélisation pour représenter cette force. Ainsi, l'énergie van der Waals distance-dependante entre deux atomes a et b peut être décrite par l'équation suivante :

$$E_{vdW} = D_o[(\frac{r_o}{r_{ab}})^{12} - (\frac{r_o}{r_{ab}})^6] \tag{1.1}$$

D_o et r_o sont des constantes, tandis que r_{ab} représente la distance entre les atomes a et b. Le terme répulsif en $(1/r_{ab})^{12}$ rend compte de façon ad-hoc d'un effet purement quantique, le principe d'exclusion de Pauli qui empêche l'interpénétration mutuelle des nuages électroniques de deux atomes. En revanche, le terme attractif dérivé des interactions de dispersion, a lui pu être démontré rigoureusement dans le cadre de la physique quantique. Le potentiel de Lennard-Jones reste à ce jour une bonne approximation des forces de van der Waals très simple à implémenter.

Si la composante répulsive du potentiel de van der Waals est nécessaire pour éviter le recouvrement des nuages électroniques en dynamique moléculaire, elle se montre beaucoup trop restrictive dans le cas du CPD. En effet l'utilisation de rotamères discrets et d'un squelette rigide entraîne inévitablement des conflits stériques qui pourraient être évités par de petits réajustement des chaînes latérales. Une fois de plus, la discrétisation de l'espace conformationnel impose des modifications empiriques de la fonction d'énergie.

Différentes stratégies furent envisagées pour pallier ce problème. La première consiste à réduire uniformément la taille des rayons de van der Waals de 5 à 10% [118], [54], [45]. L'effet du terme répulsif peut ainsi être atténué permettant alors une certaine flexibilité des chaînes latérales et du squelette peptidique. De cette façon, des rotamères initialement défavorables peuvent être repêchés. Cette approche permit l'ingénierie de protéines beaucoup plus stables en augmentant la surface hydrophobe du coeur protéique. Néanmoins elle peut introduire quelques artéfacts. Par exemple, en permettant des recouvrements atomiques, cette stratégie peut conduire à des liaisons hydrogènes artificielles, les atomes étant autorisés à se rapprocher davantage qu'avec un potentiel de Lennard-Jones classique. D'autre part, réduire les rayons de van der Waals peut entraîner un 'surpacking' du coeur hydrophobe.

Une autre approche consiste en l'agrandissement des librairies de rotamères utilisées [126], [55]. En introduisant de nouvelles orientations de chaînes latérales on décrit mieux leur flexibilité, sans pour autant résoudre le problème du squelette fixe. Cette méthode a l'avantage de ne pas altérer le potentiel de van der Waals mais elle conduit à une augmentation de l'espace conformationnel et donc du temps de calcul.

Enfin, la dernière stratégie contourne les conséquences de la discrétisation de l'espace conformationel en minimisant les rotamères individuellement. Généralement, deux étapes très brèves de minimisation sont introduites lors du calcul de la matrice d'énergie [211] [148]. La première vise à minimiser les rotamères individuels en présence du squelette, maintenu rigide. Une contrainte sur les angles dièdres est alors imposée afin de conserver le rotamère d'origine. Puis, est réalisée une seconde étape de minimisation pour chaque paire de rotamères afin de les réajuster mutuellement et ainsi limiter les conflits stériques. L'énergie après minimisation est ensuite stockée dans la matrice d'énergie. Cette méthode permet de récuperer des rotamères nécessitant de petits réarrangements sans pour autant altérer le potentiel de Lennard-Jones.

1.2.4 Liaisons hydrogènes

Les liaisons hydrogènes jouent un rôle important dans le maintien de la structure protéique, notamment dans les structures secondaires régulières telles que les hélices α et les feuillets β. Elles peuvent aussi avoir un rôle dans la réalisation de la fonction d'une protéine. La plupart des champs de force tels que CHARMM [17], AMBER [25] ou OPLS [102] traitent les liaisons hydrogènes de manière implicite à travers les énergies de van der Waals et électrostatiques. Elle peuvent cependant être représentées explicitement. C'est le cas du champ de force DREIDING implémenté par Mayo et coll. [42]. Les auteurs ont alors ajouté une contrainte sur l'orientation

des liaisons hydrogènes en introduisant un terme dépendant de l'angle entre le donneur et l'accepteur. Cette contrainte permet de favoriser la formation de liaisons hydrogènes de géométrie raisonnable. Dahiyat et coll. ont pu appliquer avec succès leur champ de force au *design* de surfaces hélicales dans des repliements *coiled-coils*. Par la suite, cette méthode fut reprise et améliorée par le groupe de Baker. Ces derniers furent capables de prédire des séquences et des interfaces protéine-protéine de type natif et réussirent d'autre part, à discriminer des complexes protéine-protéine correctement arrimés parmi un grand jeux de structures alternatives [114].

1.3 Algorithmes d'optimisation appliqués au CPD

Le CPD nécessite des algorithmes d'optimisation capables de déterminer les meilleures solutions parmi le très grand ensemble des séquences et des conformations possibles [53]. Ces algorithmes doivent faire un compromis entre la rapidité d'execution et l'exactitude des résultats. En outre, le choix de la méthode est fortement dépendant de la repésentation de l'espace conformationel et de la fonction d'energie employée.

Il existe deux types d'algorithmes d'optimisation. Les algorithmes déterministes aussi connus sous le nom d'algorithmes 'semi-exhaustifs' convergent systématiquement vers une unique solution pour un jeu de paramètres donné. Ces méthodes requièrent toutes une représentation discrète de la chaîne principale et des chaînes latérales et sont restreintes à des fonctions d'energie décomposables en une somme d'énergies de paires. Les algorithmes stochastiques ou semi-aléatoires échantillonnent de façon aléatoire l'ensemble des solutions possibles.

1.3.1 Méthodes déterministes ou semi-exhaustives

Les méthodes réellement exhaustives sont restreintes à de petits espaces conformationnels. La complexité combinatoire du CPD impose donc des méthodes semi-exhaustives qui autorisent uniquement certaines conformations discrètes afin de réduire l'espace conformationel. Deux algorithmes prédominent dans le domaine du CPD : le champ moyen et le Dead-End Elimination (DEE).

Le Champ moyen

Le champ moyen fut au départ appliqué à la prédiction de l'orientation des chaînes latérales [108] mais a ensuite vu son champ d'action s'étendre à l'ambitieux problème du CPD [112]. L'idée principale du champ moyen est de 'résumer' toutes les interactions possibles entre un rotamère donné et tous les autres rotamères de la protéine en une unique interaction moyenne. Cette méthode utilise donc une représentation où chaque rotamère interagit avec toutes les conformations possibles des chaînes latérales environnantes, ponderées par leur probabilité respective.

Le champ moyen calcule iterativement la probabilité de Boltzmann P(i,k) de chaque rotamère k pour chaque position i à partir de son énergie E(i,k) :

$$P_{i,k} = \frac{e^{-\frac{E(i,k)}{RT}}}{\sum_{l=1}^{N_i} e^{-\frac{E(i,l)}{RT}}} \tag{1.2}$$

R est la constante des gaz parfaits tandis que T est la temperature.

L'énergie E(i,k) a la forme :

$$E(i,k) = E_{BB}(i,k) + \sum_{j \neq i} \sum_{l} E(ik,jl)P(j,l) \tag{1.3}$$

Le premier terme E_{BB} représente l'energie d'interaction avec le squelette. Le second terme décrit les énergies d'interaction entre le rotamère (i,k) et tous les rotamères l pour chaque position j de la chaîne, ponderées par leur probabilité P(j,l). E(i,k) correspond donc à la moyenne des energies d'interaction entre la chaîne latérale (i,k) et son environnement. Au départ tous les rotamères d'une position donnée peuvent être considerés équiprobables. On calcule ensuite les énergies de chaque rotamère en fonction des probabilités des chaînes latérales avoisinantes (équation 1.3). De nouvelles probabilités de Boltzmann pour chaque rotamère (i,k) sont alors déduites de ces energies E(i,k) (équation 1.2).

Ce processus est répeté jusqu'à convergence des probabilités et des énergies. Afin d'éviter le phénomène d'oscillations, la convergence peut être considérablement améliorée en tenant compte d'une 'mémoire' du cycle précédent lors de la mise à jour des probabilités. Cette nouvelle condition conduit à l'expression suivante :

$$P(i,k)^{(n+1)} = \lambda P(i,k)^{(n)} + (1-\lambda)P(i,k)^{(n-1)}. \tag{1.4}$$

Le champ moyen ne garantit pas la convergence vers le minimum d'énergie global mais vers un ensemble de rotamères où chaque rotamère est le plus probable pour une position donnée. L'avantage de cet algorithme est son temps d'execution relativement rapide. Ce temps augmente linéairement avec le nombre de résidus de la protéine, permettant l'application du champ moyen à des systèmes de grande taille [109]. La figure 1.4 illustre l'ensemble du procédé.

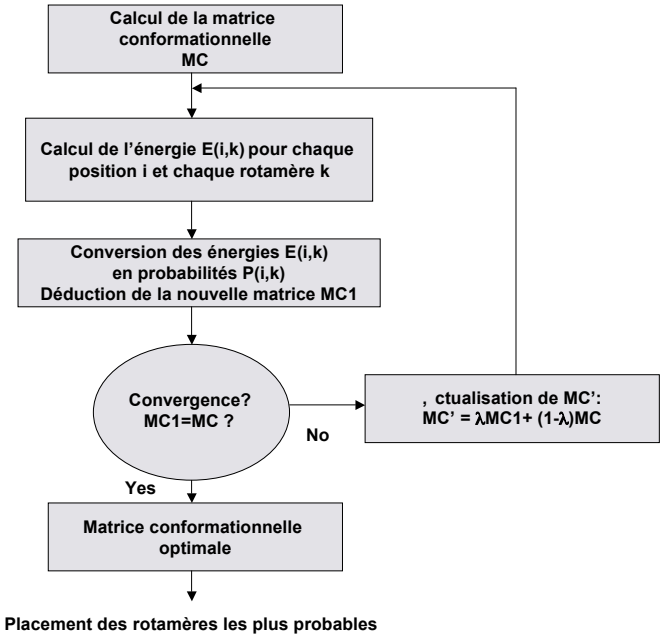


FIG. 1.4 – **Procédure générale du champ moyen.** La matrice conformationnelle MC contient l'ensemble des probabilités de chaque rotamère k pour toutes les positions i de la protéine. Les énergies E(i,k) et les probabilités P(i,k) sont décrites dans le texte.

Le Dead End Elimination

Le Dead-End Elimination (DEE) vise à déterminer la combinaison de rotamères optimal à partir d'un squelette donné en minimisant la fonction d'énergie E_{TOT}. Le principe consiste en l'élimination des mauvaises combinaisons de rotamères qui ne pourraient aboutir au minimum global[56]. De même que pour le champ moyen, cette méthode fut initialement developpée afin de prédire la structure des chaînes latérales mais a depuis largement fait ses preuves dans le domaine du CPD. L'énergie doit être décomposable en énergies de paires et doit inclure une energie 'propre' (chaîne-latérale/chaîne-principale) et une énergie 'de paires' (chaîne-latérale/chaîne-latérale).

Elle est définie par l'equation suivante :

$$E_{TOT} = \sum_{k} E_k(r_k) + \sum_{k \neq l} E_{kl}(r_k, r_l) \tag{1.5}$$

$E_k(r_k)$ représente l'énergie 'propre' d'un rotamère r_k à la position k et $E_{kl}(r_k, r_l)$ correspond à l'energie 'de paire' des rotamères r_k, r_l. Le DEE identifie les rotamères incompatibles avec le minimum global. Il existe deux critères d'élimination.

Critère simple : Elimination d'un rotamère

Un rotamère r_k^A à la position k est éliminé si la plus basse énergie obtenue avec ce rotamère est plus elevée que la plus haute énergie obtenue avec un autre rotamère r_k^B à la même position. Ce critère peut être représenté par l'expression mathématique suivante :

$$E_k(r_k^A) + \sum_{l=1}^{N} min_X E_{kl}(r_k^A, r_l^X) > E_k(r_k^B) + \sum_{l=1}^{N} max_X E_{kl}(r_k^B, r_l^X) \tag{1.6}$$

N correspond au nombre de positions de la chaîne peptidique, $minE_{kl}(r_k^A, r_l^X)$ représente la plus basse energie possible obtenue entre le rotamère r_k^A de la position k et n'importe quel rotamère X à la position l. De la même façon, $maxE_{kl}(r_k^B, r_l^X)$ est l'energie la plus elevée possible entre un rotamère r_k^B de la chaîne latérale k et n'importe quel rotamère X de la position l.

Critère double : Elimination d'une paire de rotamères

Le critère double consiste en l'élimination d'une paire de rotamères incompatible avec le minimum global. L'énergie d'une paire de rotamères A, B aux positions respectives k et l est définie comme suit :

$$U_{kl}^{AB} = E_k(r_k^A) + E_l(r_l^B) + E_{kl}(r_k^A, r_l^B) \tag{1.7}$$

La paire de rotamères A, B aux positions respectives k et l est éliminée si une autre paire C, D aux mêmes positions conduit systématiquement à une énergie plus basse. Cette condition est décrite par l'expression suivante :

$$U_{kl}^{AB} + \sum_{i=1}^{N} min_X(E_{ki}(r_k^A, r_i^X) + E_{lj}(r_l^B, r_j^X)) > U_{kl}^{CD} + \sum_{i=1}^{N} max_X(E_{ki}(r_k^C, r_i^X) + E_{lj}(r_l^D, r_j^X)) \tag{1.8}$$

Cette étape conduit par conséquent à l'élimination de la paire de rotamères A, B sans pour autant éliminer les deux rotamères individuellement. Dans certains cas, ces deux critères pèchent à éliminer certains rotamères. En effet on relève des cas où la contribution d'énergie d'un rotamère r_k^A est toujours inferieure à celle du rotamère r_k^B (pour une conformation donnée du reste de la protéine) sans que pour autant la moins bonne énergie de r_k^A soit inferieure à la meilleure de r_k^B (figure 1.5). En 1994 Goldstein introduit une variation de la méthode originale afin de pallier ce problème et permet ainsi une réduction notable du nombre de combinaisons rotamèriques [80]. Ce nouveau critère de Goldstein améliore significativement la performance du DEE en éliminant tout rotamère r_k^A pour lequel l'énergie conformationelle de la protéine en présence de r_k^A est systématiquement diminué en remplaçant uniquement r_k^A par r_k^B, le reste de la protéine étant fixé [80]. Le principe peut être illustré par la figure 1.5 et est représenté par l'expression suivante :

$$E_k(r_k^A) - E_k(r_k^B) + \sum_{l=1}^{N} min_X(E_{kl}(r_k^A, r_l^X) - E_{kl}(r_k^B, r_l^X)) > 0 \tag{1.9}$$

Le critère d'élimination de paires est l'étape limitante du DEE, réduisant son application à de petites protéines. En effet, si le critère simple augmente le temps de calcul de façon quadratique avec le nombre de résidus, le critère double l'augmente de façon cubique ! Ainsi plusieurs variantes du DEE furent implémentées en vue d'étendre leurs applications à des systèmes plus conséquents [133], [125], [137]. Par exemple, Looger et Hellinga augmentent le pouvoir de convergence de leur algorithme en considérant des jeux de rotamères plutot que des rotamères individuels.

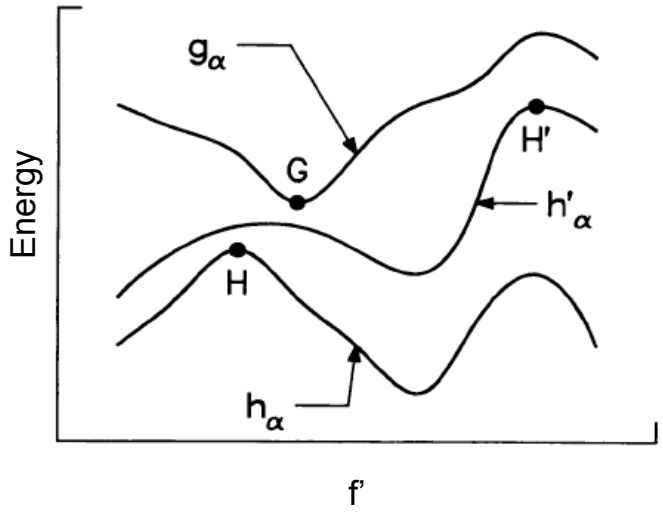


FIG. 1.5 – **Energie de la protéine en fonction des conformations f' de la protéine où la conformation du résidu α est maintenue dans la conformation g, h ou h'.** Le point G, correspondant à l'énergie la plus basse lorsque le résidu α est maintenu dans le rotamère gα, demeure plus élevé que le point H qui lui correspond à l'énergie la plus haute lorsque le résidu α est maintenu dans le rotamère hα. Ainsi, le rotamère gα est éliminé par l'équation 1.6. Cependant, si l'on considère cette fois le cas des rotamères gα et h'α, l'équation 1.6 n'éliminera pas le rotamère gα. En effet, dans ce cas le point H' correspondant à la plus haute énergie lorsque le résidu α est maintenu dans le rotamère h'α, est plus élevé en énergie que le point G. Ainsi, l'équation de Goldstein (1.9) permet d'éliminer le rotamère gα puisque la courbe d'énergie de ce rotamère est en tout point supérieure à la courbe d'énergie de h'α. Figure extraite de l'article de Goldstein [80].

Nous venons de présenter les deux algorithmes principaux parmis les méthodes déterministes utilisées en CPD. Bien que ces deux méthodes convergent chacune vers une solution unique pour un jeu de paramètres donné, seul le DEE garantit la convergence vers le minimum global. Le DEE semble donc être actuellement la méthode la plus puissante pour déterminer la meilleure combinaison rotamèrique [206]. Néanmoins, la recherche exhaustive des solutions rend le DEE très coûteux et les méthodes stochastiques restent pour l'instant les plus appropriées pour résoudre les problèmes de complexité combinatoire plus importante.

1.3.2 Méthodes stochastiques ou semi-aléatoire

Les méthodes stochastiques, échantillonnent aléatoirement l'espace des séquences et des structures en se déplaçant d'une solution à l'autre d'une façon dépendante du paysage énergétique et des lois imposées par l'algorithme d'optimisation. Plus faciles à implémenter, les méthodes stochastiques sont aussi beaucoup plus rapides que les méthodes exhaustives. Les principaux algorithmes utilisés en CPD sont le Monte Carlo [152] et l'Algorithme Génétique [91], cependant nous détaillerons aussi l'algorithme heuristique implémenté par l'équipe de Wodak [211].

Monte Carlo

Le Monte Carlo est l'une des méthodes stochastiques les plus simples à implémenter. Son principe est de proposer itérativement une modification au modèle étudié puis de décider d'accepter ou rejeter cette modification selon le critère de Metropolis. Cette méthode peut être utilisée pour optimiser des séquences d'acides aminés, des orientations de chaînes latérales, des conformations de chaînes principales ou encore tous ces critères simultanément [87],[43], [78], [92].

Dans le cas du CPD, un acide aminé dans un rotamère donné est aléatoirement modifié pour une position choisie au hasard dans toute la séquence. La nouvelle énergie du système E_{new} est alors mise à jour. Si cette énergie est plus faible que l'énergie E_{old} de l'état précédent alors la modification est acceptée. Si au contraire cette énergie se trouve plus élevée, la perturbation est acceptée avec la probabilité $p = exp^{(E_{new}-E_{old})/kT}$. k représente la constante de Boltzmann et T la température. Cette opération est répétée un grand nombre de fois. La méthode de Monte Carlo permet de franchir les barrières d'énergies et ainsi surmonter les multiples minima locaux du paysage énergétique.

La température peut être ajustée pour faciliter le franchissement de barrières d'énergie. Le recuit simulé reprend ce principe en chauffant le système puis en le refroidissant afin de diminuer graduellement la probabilité d'accepter des conformations de haute énergie.

Algorithme heuristique de Wernisch et coll.

En 2000, Wernisch et coll. implémentèrent un algorithme permettant l'identification de la séquence de plus basse énergie mais aussi la collection d'un ensemble de séquences d'énergie similaires [211]. Dorénavant, pour plus de lisibilité, nous désignerons les combinaisons structures/séquences de basse énergie par le terme de 'séquences de basse énergie'. Cette approche repose sur le fait que le paysage énergétique décrit par nos champs de force ne représente pas le vrai paysage énergétique. Par conséquent, les énergies libres obtenues lors des calculs de CPD demeurent une approximation des énergies réelles. Ainsi, la séquence optimale obtenue ne correspond pas nécessairement à la meilleure solution. Wernisch et coll. ont pris pour parti de collecter un ensemble de solutions d'énergie proches de la plus basse énergie plutôt que de se limiter à une seule solution qui ne correspond probablement pas au minimum global réel.

L'algorithme utilise un ajustement itératif des rotamères à chaque position. Au départ, un rotamère est imposé aléatoirement pour chaque position. Ensuite une position i est choisie au hasard et le rotamère optimal pour cette position est déterminé, compte tenu des autres rotamères aux autres positions. Le meilleur rotamère est alors assigné à la position i et une nouvelle position j est ensuite choisie au hasard. De nouveau, le meilleur rotamère pour cette position est déterminé en tenant compte des chaînes latérales environnantes. La procédure est répétée jusqu'à convergence de l'énergie.

Cet algorithme, tout comme le Monte Carlo, ne garantit pas de trouver le minimum global mais un minimum local proche du point de départ. Toutefois, en multipliant les points de départ et en répétant les cycles, Wernisch et coll. montrent que le minimum global peut facilement être atteint pour des petits systèmes autour de 20 résidus dans des temps beaucoup plus rapides que le DEE [211].

Algorithme Génétique

L'algorithme génétique fut developpé par Holland et ses collègues en 1975 et est largement exploité en prédiction stucturale ou en CPD depuis le début des années 90 [201], [101]. Cette approche a pour but d'optimiser une population de solutions en s'inspirant d'opérations biologiques tels que la mutation, la recombinaison et la sélection. Les deux premiers sont des opérations d'exploration de l'espace, tandis que le dernier fait évoluer la population vers les optima du problème.

Une population k de solutions est tout d'abord générée au hasard. Des mutations aléatoires sont alors appliquées avec un taux défini au préalable. Les énergies des mutants sont ensuite mises à jour, puis les différentes solutions sont classées selon leurs énergies. Opère alors le processus de pression de sélection qui va identifier un jeu de solutions de basse énergie. Ensuite, on recombine les solutions de basse énergie entre elles afin de privilégier la reproduction des 'meilleurs' individus au détriment des moins 'bons'. Enfin, les meilleures séquences d'un point de vue énergétique sont gardées pour peupler la nouvelle population $k + 1$. Le nombre de solutions à conserver d'un cycle à l'autre peut varier selon les différentes approches. En prenant un nombre trop faible on prend le risque de converger trop rapidement vers des minimas locaux et de se restreindre à une trop petite région de l'espace. A l'inverse, garder trop de solutions ferait augmenter le temps de calcul. Une des solutions communément adoptée est de garder constant le nombre d'individus d'une

génération à l'autre. Une fois la nouvelle population obtenue, ce processus est réitéré jusqu'à l'obtention d'une population homogène dont on peut penser qu'elle se situe à proximité du ou des optima.

Comme dans le Monte Carlo, il existe certaines variantes de l'algorithme génétique analogues au recuit simulé. Cette fois, c'est le nombre d'individus conservés d'une génération à l'autre qui décroit au cours des différents cycles. Le principe est de commencer avec une large distribution de solutions dans l'espace des séquences puis de diminuer le nombre de solutions à chaque cycle jusqu'à converger vers une solution unique. On espère ainsi augmenter la probabilité de trouver de meilleurs minima.

Un des avantages de l'algorithme génétique est qu'il permet de franchir des barrières d'énergies par des déplacements dans l'espace des séquences, d'amplitude beaucoup plus grande que ceux autorisés par l'algorithme de Monte Carlo. Ceci peut cependant se révéler être un inconvénient pour les systèmes fortement couplés. Par exemple, le mécanisme de recombinaison peut s'avérer problématique dans le cas de la prédiction de chaînes latérales puisque de nombreux résidus proches dans la séquence ne le sont pas nécessairement dans la structure et inversement.

Actuellement les algorithmes stochastiques ne garantissent pas la détermination du minimum global ni même le fait d'explorer des solutions proches de ce minimum. Cependant ils ont largement démontré leurs capacités dans le *design* de nombreux coeurs hydrophobes [54], [126], [87] Voigt et coll. confirment que le DEE reste actuellement l'algorithme le plus approprié pour converger vers le minimum global [206]. On peut cependant relativiser l'importance de déterminer le minimum global, l'utilité d'une réponse très précise n'étant pas évidente. En effet, il faut tout d'abord distinguer le minimum global de l'espace conformationel discret utilisé dans toutes ces approches, du minimum global réel. Le paysage énergétique, décrit par

nos champs de force, ne représente pas le vrai paysage énergétique. Par conséquent, la nécessité de trouver le minimum global dans le problème du CPD n'est pas absolue, voire même erronée. Rechercher un ensemble de solutions proches du minimum global théorique semble donc plus pertinent.

1.4 Calibration de la procédure de CPD

Le CPD requiert une paramétrisation très fine des différents termes de la fonction d'énergie. On dénombre différentes approches pour calibrer et pondérer ces composantes énergétiques, reposant soit sur des critères statistiques soit sur des critères physiques.

Dans une approche statistique, Kuhlman et Baker ont optimisé leur fonction d'énergie de façon à génerer des séquences de type natif pour un jeu test de protéines [118]. Pour ce faire, ils essaient itérativement les 20 acides aminés à chaque position de la protéine étudiée tout en maintenant fixées les autres chaînes latérales. Les différents termes énergétiques sont alors pondérés de façon à maximiser la probabilité de Boltzmann d'obtenir le résidu natif pour chacune des positions. Ce protocole fut calibré sur un jeu d'une trentaine de protéines puis appliqué en 'aveugle' à une centaine de protéines. Ils obtiennent ainsi des séquences très proches de la native où 51% des résidus du coeur et 27% de tous les résidus prédits par leur algorithme sont strictement identiques aux résidus natifs dans les positions correspondantes. Toutefois, cette approche pourrait introduire un biais dans la fonction d'énergie. En effet, elle s'appuie sur l'hypothèse que la séquence native serait la plus stable énergétiquement. Or, les séquences naturelles ne sont pas nécessairement optimisées pour assurer la stabilité maximale mais pour satisfaire un compromis entre la stabilité et la réalisation de la fonction protéique [189].

Une autre approche statistique consiste à maximiser l'écart d'énergie entre la séquence native et un ensemble de séquences aléatoires enfilées sur la même structure. Pratiquement, ceci revient à maximiser le Z-score, c'est à dire, l'écart d'énergie entre la séquence native cible et la moyenne de la distribution énergétique des séquences aléatoires normalisée par l'écart type de la distribution [34]. Street et coll. ont repris cette méthode pour le *design* de feuillets β de la protéine G de streptocoque et un variant de l'apoplastocyanine [192]. Ils réussirent ainsi à prédire des séquences conduisant à des structures aussi stables voire plus stables que les structures natives.

Enfin, une approche fondée sur des critères physiques est la prédiction de changements de stabilité associés à des mutations ponctuelles. La fonction d'énergie est optimisée de façon à réduire l'écart entre les énergies prédites et celles obtenues expérimentalement. Cette approche n'impose aucune contrainte sur la séquence et a conduit à des résultats très satisfaisants [211], [43] [126], [55].

Il existe donc plusieurs méthodes de CPD qui diffèrent par la méthode d'optimisation de séquences, la définition de la fonction d'énergie, ou encore sa paramétrisation. Une étude réalisée par le groupe de Farid montre qu'avec leur algorithme de CPD, les séquences obtenues pour un jeu de protéines ressemblent davantage aux séquences prédites par les autres groupes qu'aux séquences natives [100]. Ceci suggère qu'en dépit des nombreuses approches utilisées, les algorithmes de CPD tendent vers des résultats globalement similaires.

1.5 Champ d'application du CPD

1.5.1 Premiers succès du CPD

Design de coeurs hydrophobes

Les premières applications du CPD furent concentrées sur le *design* de coeurs hydrophobes puisque ces derniers constituent la région la plus simple, les interactions van der Waals étant prédominantes dans le maintien de la structure de ces régions. Desjarlais et Handel réalisèrent l'ingénierie de différents variants de la protéine 434 cro en mutant 5, 7 et 8 positions du coeur hydrophobe. Au final, deux des variants sur trois présentaient une stabilité supérieure à celle de la protéine native [54]. Dahiyat et Mayo mirent en place une procédure automatique de CPD dont la fonction d'énergie reposait essentiellement sur un potentiel de van der Waals. Leur algorithme fut testé dans sa capacité à prédire des séquences de type natif à partir du coeur hydrophobe d'un homodimère *coiled-coil* extrait de la protéine de GCN4. Malgré l'utilisation d'un champ de force très simple, les séquences obtenues se repliaient dans des structures stables présentant des caractéristiques expérimentales très proches de celles de la protéine native [43] . On compte bien d'autres succès dans le *design* de coeurs hydrophobes tels que l'ingénierie de variants de coeurs d'ubiquitine par Lazar et coll. ou encore le *design* du coeur du domaine β1 de la protéine G de streptocoque réalisé par Dahiyat et Mayo [45], [126]. Ces travaux pionniers constituaient à l'époque un premier défi dans le domaine du CPD et représentaient une bonne évaluation des différentes fonctions d'énergies implémentées jusqu'alors.

Design de surfaces protéiques

Si le coeur des protéines peut être décrit par des potentiels d'énergie très simples, ceci est loin d'être le cas pour les surfaces protéiques. En effet, les surfaces protéiques autorisent une grande diversité de séquence, augmentant donc le nombre de degrés de

liberté. Par ailleurs, dans ces régions, les interactions électrostatiques prédominent sur celles de van der Waals. De plus, les résidus de surface sont contraints par la nécessité d'accomplir une fonction protéique via des interactions protéine-protéine ou protéine-ligand. Enfin, en surface les interactions chaîne latérale-solvant priment sur les interactions chaîne latérale-chaîne latérale. Or, ces interactions se révèlent difficiles à modéliser dans la mesure où le CPD requiert l'utilisation d'un solvant implicite. La modélisation des différents phénomènes survenant à la surface des protéines impose donc une fonction d'énergie plus fine et plus complexe. Néanmoins, on relève quelques tentatives réussies de *design* de surfaces protéiques [192].

Design **de protéines entières**

Actuellement, on compte peu d'exemples de *design* de protéines entières. Dahiyat et Mayo furent les premiers à relever ce défi, avec l'ingénierie d'un motif $\beta\beta\alpha$ de 28 résidus structurés en doigts de zinc [44], [46]. Aucune contrainte sur la séquence n'avait été ajoutée ce qui était novateur pour l'époque. La séquence obtenue présentait un score d'identité de 21 % avec la séquence native. Les auteurs constatèrent sans surprise que 75% de ces identités étaient localisées dans le coeur, les résidus de surfaces se révélant beaucoup moins bien conservés. Malgré ce faible score d'identité, leur séquence fut capable de se replier dans une structure compacte et stable très similaire à la structure native. Après ce succès, d'autres prouesses similaires se succédèrent telles que le *design* complet de différents variants d'homéodomaines [140] ou de la séquence complète du domaine WW [115].

Cependant ce type d'étude n'est pas très répandu. En effet, selon les applications, redessiner la protéine dans sa totalité peut se révéler utile ou au contraire infructueux et coûteux. On note toutefois quelques applications telle que la reconnaissance de pli nécessitant le *design* de la protéine dans son intégralité. Ceci sera détaillé en section 1.5.3.

1.5.2 Exploration de l'espace de repliement des protéines

Les nombreuses avancées dans le domaine du CPD ont ouvert de nouveaux horizons notamment sur l'exploration du paysage énergétique de repliement des protéines[119], [82]. Les récents progrès observés dans ce domaine nous permettent de poser des questions de plus en plus ambitieuses. Les protéines ont-elles été optimisées par la sélection naturelle pour se replier rapidement ? En quelle mesure la nature a t-elle échantillonné toutes les régions de l'espace conformationnel des protéines ? Dans une protéine donnée, quelles sont les caractéristiques qui découlent directement de la sélection naturelle ou des propriétés physiques fondamentales des protéines ? Différentes applications ont ainsi vu le jour afin de proposer des éléments de réponses à ces nombreuses questions.

Stabilisation de protéines

L'étude de la stabilité des protéines fut l'une des premières applications du CPD. D'une part, cela nous permet d'élargir nos connaissances sur les forces déterminant la structure et la stabilité des protéines. D'autre part, le *design* de protéines thermostables présente un intérêt industriel important.

On rapporte de nombreux exemples d'ingénierie de protéines présentant des stabilités accrues par rapport à la protéine native [139], [68], [47]. De façon générale toutes ces études constatent que la manière la plus simple pour stabiliser une protéine est d'augmenter la proportion de résidus hydrophobes du coeur protéique. Cette approche rencontra toutefois ses limites au cours d'un travail très intéressant du groupe de Mayo. Ces derniers échouèrent dans la tentative de stabiliastion du lysozyme de T4 suggérant que la séquence du lysozyme était proche de la séquence optimale pour cette structure [153].

Le groupe de Serrano échoua aussi dans la tentative d'augmenter la stabilité de domaines SH3 de spectrine en créant un phénomène de sur-compacité du coeur hydrophobe [202]. Les mutations prédites conduisaient à un volume de résidus hydrophobes enfouis supérieure à celui observé dans la protéine native, résultant en une tension au sein du coeur protéique. Il est intéressant de noter que cette tension pouvait être levée par la suppression d'un groupe méthyl réduisant ainsi le volume du coeur hydrophobe. Les nouveaux mutants retrouvaient la stabilité et les propriétés structurales de la protéine native.

Ces résultats montrent que le remplacement systématique des résidus polaires enfouis par des résidus non polaires peut s'avérer 'dangereux' pour le maintient de la stabilité d'une protéine. A cet égard, Mayo et coll. ont développé récemment, un filtre basé sur le nombre de liaisons hydrogènes formées par les différents résidus du coeur afin de déterminer les résidus polaires devant être maintenus intacts lors du *design* de coeurs protéiques [14].

Modification de chemins de repliement

En 1983, Go implémenta un modèle pour décrire les mécanismes du repliement des protéines. Ce modèle, repose sur la connaissance de la structure native et suppose que les interactions non natives ne jouent pas de rôle dans le processus de repliement. Au contraire ce modèle fait l'hypothèse que le chemin de repliement est celui qui maximise la formation d'interactions natives tout en minimisant la perte d'entropie de la chaîne principale [77]. Une autre théorie décrit le paysage énergétique du processus de repliement comme un entonnoir dans lequel le repliement natif et les structures quasi natives occupent le bassin de plus basse énergie [22], [103]. Aujourd'hui, les nouveaux algorithmes s'appuient sur une combinaison de ces deux modèles

en considérant uniquement les interactions natives et montrent de bonnes performances dans la prédiction des principaux événements survenant pendant le processus de repliement. En particulier, on relève de nombreuses méthodes de prédictions des régions impliquées dans le repliement [155], [3], [72]. Ceci devient alors fructueux pour optimiser des chemins de repliements, modifier l'ordre des événements survenant pendant celui-ci, ou encore concevoir de nouveaux chemins de repliements.

Plusieurs groupes s'attelèrent à l'accélération du processus de repliement [156], [203], [33]. Le principe consiste à stabiliser les régions de la protéine déjà structurées pendant l'étape limitante de la réaction qui correspond à la formation du noyau de repliement. Serrano et coll. mutèrent différents résidus de la procarboxypeptidase humaine afin de stabiliser localement deux hélices α de son domaine d'activation [203]. Les multiples mutations de la seconde hélice résultèrent en une accélération du processus de repliement et un ralentissement du dépliement de la protéine mutante. D'autres résultats similaires furent observés lors d'une étude effectuée sur l'acylphosphatase [196]. Ces travaux montrent qu'introduire des interactions favorables dans des régions intervenant directement dans les premiers événements du repliement peut conduire à une accélération de ce processus. Cette découverte suggére que les protéines n'ont pas été naturellement optimisées pour maximiser leur rapidité de repliement.

Le cas des fibres amyloides présente un intérêt particulier. Différentes études ont montré que les intermédiaires (partiellement structurés) du repliement pouvaient jouer un rôle précurseur dans le processus d'agrégation et la formation de fibres amyloïdes [171], [204]. L'idée serait alors de façonner le chemin de repliement en introduisant des contraintes dans le paysage énergétique afin d'éviter ces intermédiaires. Pratiquement, cela consisterait à défavoriser les chemins non-natifs et stabiliser les régions impliquées dans le noyau de repliement. Merkel et Regan

testèrent cette hypothèse sur la Green Fluorescent Protein avec succès, créant des mutants se repliant plus rapidement que la protéine native et moins propices au phénomène d'agrégation [151].

Dans un registre légèrement différent, le groupe de Baker présente un exemple remarquable de modification du chemin de repliement des protéines symétriques G et L par de simples méthodes de CPD [144], [106]. Ces protéines sont constituées d'une hélice α et de deux épingles à cheveux β, la seconde épingle à cheveux β étant formée en premier dans le cas de la protéine G tandis qu'elle apparaît en dernier dans le cas de la protéine L. Or en stabilisant dans chacune des protéines l'épingle à cheveux β formée en dernier et en destabilisant l'épingle à cheveux opposée les auteurs réussirent à inverser les chemins de repliement respectifs des protéines G et L. Cette étude est une belle démonstration de la progression de notre compréhension des différents mécanismes déterminants du repliement protéique.

Design de nouveaux repliements structuraux

Bien qu'il existe un grand nombre de repliements protéiques différents, ce nombre n'est pas infini. Que certains repliements n'aient jamais été observés peut signifier qu'ils ne sont pas réalisables d'un point de vue physico-chimique, qu'ils n'ont pas été encore échantillonnés par le processus d'évolution ou qu'ils n'ont tout simplement pas encore été identifiés par l'homme. Une des tentatives les plus spectaculaires pour répondre à cette question fût entreprise par le groupe de Baker [120]. Ils mirent en place un cycle itératif de sélection de séquences et optimisation de squelettes peptidiques afin de créer un nouveau repliement α/β encore jamais observé dans la nature (figure 1.6).

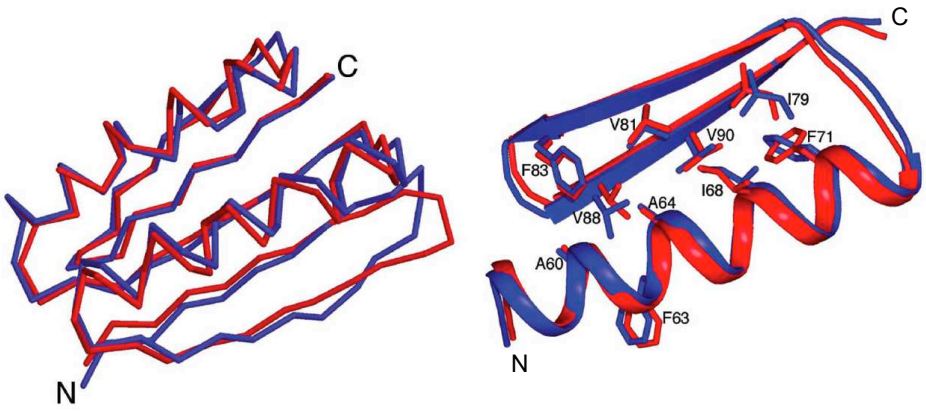


FIG. 1.6 – **Comparaison du modèle 3D (bleu) et de la structure cristallographique (rouge) de Top7.** A gauche représentation de la chaîne principale. A droite représentation du squelette peptidique en mode 'cartoon'. Les chaînes latérales du coeur protéiques sont représentées en mode 'stick'. Figure extraite de l'article de Kuhlman et coll. [120].

Leur modèle, Top7 fut validé expérimentalement par des analyses de dichroïsme circulaire et de dénaturation chimique révélant une structure très stable. En effet, Top7 présentait une stabilité de 13,2 kcal/mol. Une fois de plus, ce résultat suggère que les protéines sont loins de l'optimum thermodynamique. Par ailleurs, la structure de Top7 présentait un parfait accord avec le modèle prédit, le rmsd entre ce dernier et la structure expérimentale étant de 1.2 Å. Ainsi, cette étude montre que l'absence de ce repliement dans la Protein Data Bank[2] ne provient pas de raisons physico-chimiques. Cette méthode s'avère très prometteuse dans l'exploration d'un nouvel espace de structures et architectures protéiques, ne limitant plus nos futures recherches à l'espace des structures déjà échantillonnées par le processus de l'évolution naturelle.

[2]www.rcsb.org

1.5.3 Application génomique : reconnaissance de repliement

Depuis quelques années, le CPD commence à jouer un rôle important dans le domaine de la prédiction structurale [191]. Une méthode originale de reconnaissance de plis repose sur la résolution du problème inverse du repliement. Le principe consiste à rechercher la séquence optimale pour un repliement donné au lieu de rechercher la structure optimale pour une séquence donnée. Ainsi, plutôt que d'explorer l'immense espace des conformations, cette approche permet de réduire le domaine de recherche à l'espace des séquences [110], [118], [113], [170].

La modélisation par homologie repose sur l'hypothèse suivante : si une séquence de structure inconnue est similaire à une séquence de structure connue on peut supposer que la première adopte le même repliement que la seconde. En pratique, à partir d'une séquence de structure inconnue, on recherche les séquences qui lui sont homologues et dont la structure est connue. Pour ce faire, on réalise des alignements de séquences entre la séquence d'intérêt et chacune des séquences pour lesquelles la structure a été résolue. On suppose alors que la séquence d'intérêt se replie de la même façon que ses séquences homologues. Généralement, on attribue pour les régions conservées la même conformation pour la chaîne principale. Les boucles additionnelles peuvent être prédites par des méthodes de modélisation 'ab initio'. Le positionnement des chaînes latérales est déterminé à l'aide des algorithmes d'optimisation mentionnés en section 1.3. Enfin, le modèle est affiné par des méthodes de minimisation d'énergie ou de dynamique moléculaire. La modélisation par homologie est d'autant plus efficace que le nombre de séquences adoptant un repliement donné est grand puisque cela permettra d'identifier plus facilement les homologues de la séquence de départ. C'est pourquoi, le CPD est appliqué à grande échelle dans le but d'identifier toutes les séquences compatibles avec une structure particulière, définissant ainsi une 'signature' de ce repliement.

1.5.4 *Design* de nouvelles fonctions biologiques

Ces dernières années, le champ d'application du CPD s'est déplacé vers le *design* de nouvelles fonctions [127]. La réalisation des différentes fonctions biologiques requiert généralement l'association de deux protéines, d'une protéine et d'un acide nucléique ou encore d'une protéine et son ligand. On rapporte diverses approches selon la nature de la fonction et donc de l'interaction désirée. Marvin et Hellinga réalisèrent la conversion d'un récepteur au maltose en un biosenseur capable de lier le zinc [143]. Dans une autre étude, le groupe de Hellinga introduisit une activité triose phosphate isomérase dans une protéine normalement dépourvue d'activité enzymatique [60]. Partant de la Ribose Binding Protein, récepteur périplasmique sans aucune activité catalytique connue, ils mutèrent les résidus du site de liaison du ribose pour obtenir une protéine biologiquement active capable d'interconvertir le dihydroxyacétone phosphate (DHAP) en glycéraldéhyde tri-phosphate (GAP).

Le groupe de Baker en collaboration avec les groupes expérimentalistes de Stoddard et Monnat réalisèrent une étude remarquable en créant une endonucléase artificielle [31]. Pour ce faire, ils fusionnèrent deux domaines d'endonucléase n'interagissant pas naturellement entre eux, via le *design* d'une nouvelle interface entre ces deux protéines. La protéine résultante montra alors une bonne stabilité ainsi qu'une bonne affinité pour son ADN cible : un long brin chimérique contenant les deux sites de liaisons aux domaines de départ.

Le CPD recouvre ainsi de nombreuses approches. Leur diversité provient principalement du fait que la fonction d'énergie requiert des ajustements empiriques qui diffèrent d'un laboratoire à l'autre. D'autre part, l'optimisation de la fonction d'énergie dépend du type d'application réalisée. En particulier, nous avons vu que cette optimisation pouvait être fondée sur des critères physiques ou statistiques. Dans le cadre de notre étude, nous nous sommes intéressés à l'évolution dirigée de deux aminoacyl-ARNt synthétases : l'aspartyl-ARNt synthétase (AspRS) et l'asparaginyl-ARNt synthétase (AsnRS). Dans le chapitre suivant, nous présentons le rôle et l'organisation structurale de ces protéines.

Chapitre 2

Les aminoacyl-ARNt Synthétases

2.1 Rôle biologique des aminoacyl-ARNt Synthétases

2.1.1 La synthèse protéique

La biosynthèse des protéines consiste en la traduction de l'information génétique portée par l'ARN messager (ARNm) en une chaîne protéique. Cette étape est assurée par la machinerie de synthèse comprenant notamment les ribosomes, les ARN de transfert (ARNt), et les aminoacyl-ARNt Synthétases (aaRS). Le ribosome est constitué d'ARN ribosomique et de protéines et comprend deux sous-unités. Au départ, la petite sous unité du ribosome se fixe sur l'ARNm. La grande sous unité assure la synthèse de la liaison peptidique entre les différents acides aminés consécutifs qui constitueront la future protéine. Pour ce faire, le ribosome parcoure le brin d'ARNm codon par codon et va ajouter un acide aminé à la protéine en cours de fabrication. Chaque acide aminé est apporté par un ARN de transfert (ARNt) qui lui est spécifique. Ainsi, les ARNt jouent le rôle d'adaptateurs entre le monde des acides nucléiques et celui des protéines (figure 2.1). L'ARNt se fixe à l'ARNm par l'intermédiaire de son anticodon (triplet de bases d'ARN) complémentaire du codon porté par l'ARNm, permettant ainsi la jonction physique entre le monde des acides

nucléiques et celui des acides aminés. L'appariement codon-anticodon a lieu dans le ribosome qui vérifie la complémentarité des bases de l'ARNm et l'ARNt. Enfin, le ribosome catalyse la liaison peptidique entre l'acide aminé porté par l'ARNt et le dernier acide aminé incorporé dans la protéine en cours de synthèse. Dès que la chaîne d'acides aminés est terminée, elle se détache du ribosome pour être libérée dans l'organisme. Le ribosome se détache à son tour de l'ARNm sur lequel il était fixé et redevient disponible pour une nouvelle synthèse protéique. L'ensemble de ce processus est illustré par la figure 2.2. La fiabilité de la traduction est garantie par les ARNt qui assurent la correspondance entre l'information génétique portée par l'ARNm et l'acide aminé incorporé dans la chaîne protéique. Cette bonne correspondance dépend directement de l'étape d'aminoacylation.

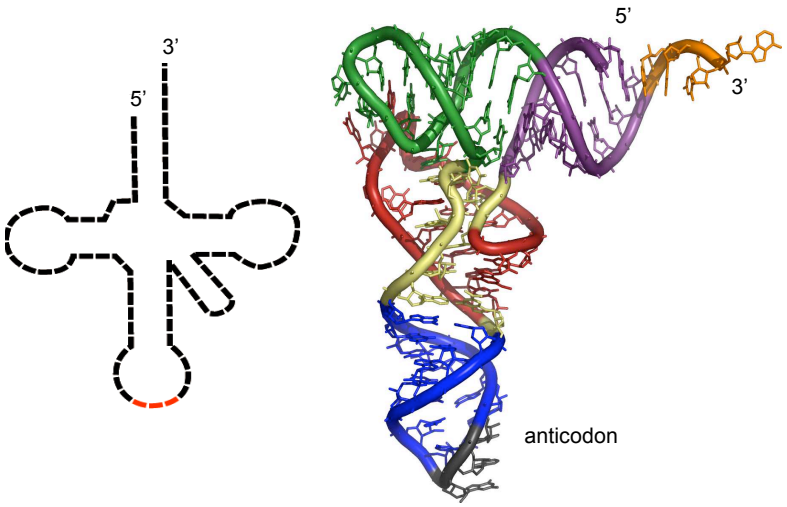


FIG. 2.1 – **Représentations 2D (à gauche) et 3D (à droite) d'un ARNt.** A gauche, l'extrémité 3'OH sur laquelle se fixe son acide aminé spécifique est indiquée. A droite, la zone de fixation de l'acide aminé est représentée en violet. La partie orange correspond à la zone de reconnaissance de l'ARNt par les synthétases. Enfin, la région représentée en noir correspond à l'anticodon qui va reconnaître le triplet de bases d'ARN porté par l'ARNm qui lui est complémentaire. Ceci va ainsi permettre la liaison de l'ARNt à l'ARNm.

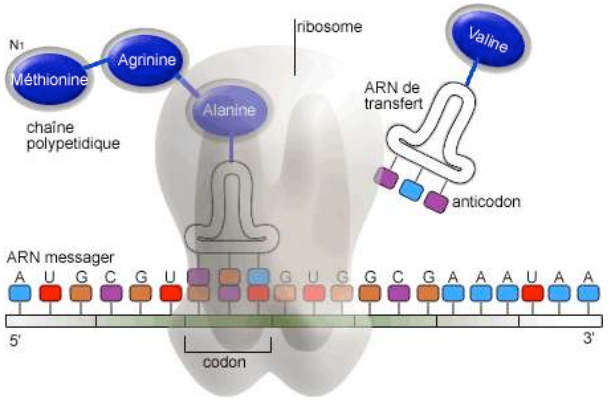


FIG. 2.2 – **Représentation schématique de la synthèse protéique.**

2.1.2 L'aminoacylation des ARNt

Il existe dans chaque organisme une aaRS spécifique d'un type d'acide aminé et de son ARNt correspondant. Leur rôle principal consiste à estérifier son acide aminé spécifique à l'extrémité de son ARNt correspondant. Cette réaction d'aminoacylation est catalysée par l'aaRS, par un mécanisme en deux étapes. L'énergie nécessaire pour l'ensemble de la réaction provient de l'hydrolyse de l'ATP [184], [123]. La première étape consiste en l'activation de l'acide aminé pour former un aminoacyl-adénylate (figure 2.3), la seconde, correspond à son transfert sur son ARNt. La réaction est décrite par l'équation chimique suivante :

$$aa + ATP \rightleftarrows aaAMP + PP_i \tag{2.1}$$

$$aaAMP + ARNt \rightleftarrows aaARNt + AMP \tag{2.2}$$

L'activation de l'acide aminé se déroule donc en présence d'ATP. Cette réaction conduit à la libération d'un pyrophosphate. Dès lors, l'acide aminé activé

peut être transféré sur l'hydroxyle 2' ou 3' de l'adénosine terminale de son ARNt spécifique.

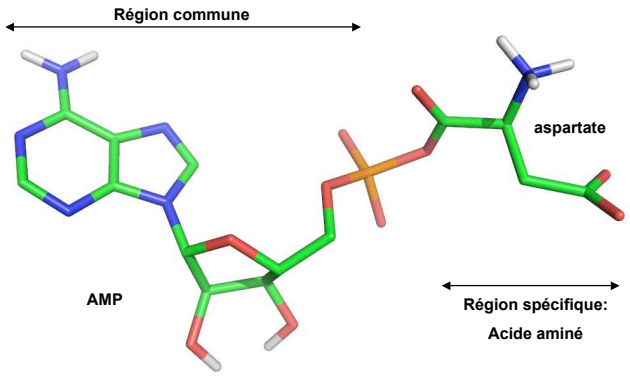


FIG. 2.3 – **Représentation de l'aspartyl-adénylate.** La région AMP est commune à toutes les synthétases, la partie spécifique correspond à l'acide aminé.

L'aminoacylation constitue l'une des étapes clés de la traduction. Des erreurs survenant à cette étape peuvent entrainer l'incorporation erronée d'acides aminés dans la protéine, puisqu'elle constitue un des derniers contrôles avant la catalyse de la liaison peptidique entre le nouvel acide aminé et la chaîne protéique [30]. En effet, le ribosome contrôle uniquement la bonne complémentarité entre le codon et l'anticodon qui est indépendante de la nature de l'acide aminé porté par l'ARNt. Les synthétases doivent donc être capables de discriminer leur acide aminé spécifique parmi les 20 acides aminés naturels, ainsi que leurs analogues ou dérivés. En outre, elles doivent sélectionner le bon ARNt parmi la soixantaine d'ARNt différents. Cette discrimination s'avère difficile compte-tenu, de l'analogie structurale existant entre certains acides aminés et de la structure en 'L' commune à tous les ARNt. Une fonction d'édition fut donc développée par les synthétases afin de corriger les erreurs d'aminoacylation. Cette étape correspond à l'hydrolyse de la liaison entre l'ARNt et le mauvais acide aminé.

2.2 Structure des aaRS

Les aaRS sont modulaires. Chacune d'entre elles, présente un coeur de 300 à 400 résidus correspondant au domaine du site actif. Peuvent ensuite s'ajouter ou s'insérer d'autres modules peptidiques dont la fonction consiste dans la plupart des cas à la reconnaissance et la liaison de l'ARNt. Ils sont apparus au cours du processus d'évolution et sont en partie responsables de l'actuelle diversité observée au sein des aaRS. De nombreuses études tentèrent en vain d'unifier la famille des aaRS en cherchant des corrélations structurales ou fonctionnelles. C'est finalement le groupe de Moras qui révéla à travers des analyses de séquences certaines homologies au niveau de leur site actif permettant de les regrouper en deux classes de 10 membres chacune [65]. Ces classes diffèrent par leurs structures, leur mode de fixation à l'ARNt et leur mécanisme d'aminoacylation, (figure 2.4) [39], [51], [49], [50]. Leur répartition est illustrée par la table 2.2.

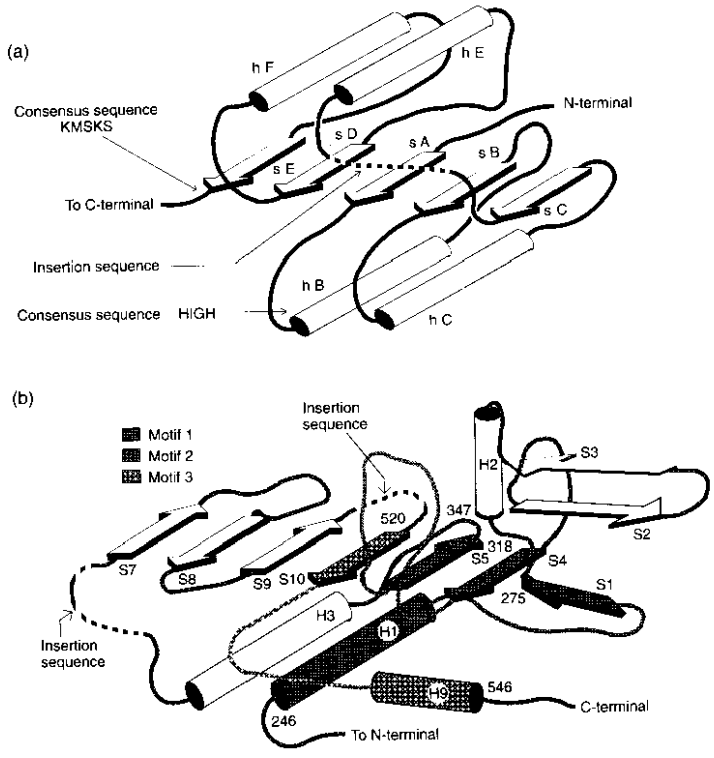


FIG. 2.4 – **Représentation schématique du site actif des aminoacyl-ARNt-synthétases de classe I (a) et classe II (b).** Figure extraite de l'article d'Arnez et Moras [7].

Sous Classe	Classe I	Str. Quaternaire	Classe II	Str. Quaternaire
	Leu	α	His	$\alpha 2$
	Ile	α	Pro	$\alpha 2$
a	Val	α	Ser	$\alpha 2$
	Cys	α	Thr	$\alpha 2$
	Met	$\alpha 2$		
	Glu	α	Asp	$\alpha 2$
b	Gln	α	Asn	$\alpha 2$
	Arg	α	Lys	$\alpha 2$
	Tyr	$\alpha 2$	Gly	$\alpha 2 \beta 2$
c	Trp	$\alpha 2$	Ala	$\alpha 4$
			Phe	$\alpha 2 \beta 2$
	2'OH		3'OH	

Table 2.1 : Répartition des aminoacyl-ARNt synthétases. La table présente la répartition des aminoacyl-ARN-t synthétases ainsi que leur structure oligomérique. Les aaRS peuvent attacher leur acide aminé respectif au niveau du 2'OH (classe I) ou du 3'OH (classe II à l'exception de la PheRS) de l'adénosine terminale de l'ARNt.

2.2.1 La classe I

Les synthétases de classe I sont monomèriques ou dimèriques. Les aaRS de cette classe sont caractérisées structuralement par un domaine de Rossman lui même constitué de 5 brins β parallèles associés par des hélices α (figure 2.4), [176], [16], [177], [210].

Par ailleurs, la plupart des aaRS de classe I comprennent les deux peptides signatures HIGH (His-Ile-Gly-His) et KMSKS (Lys-Met-Ser-Lys-Ser) ou des séquences analogues, localisés dans le site actif [65], [94], [23], (figure 2.5). Ces deux motifs conservés interagissent directement avec l'ATP qui adopte une conformation étendue. Par ailleurs, différentes études leur attribuent un rôle dans la stabilisation de l'état de transition de l'adénylate [129], [21], [187].

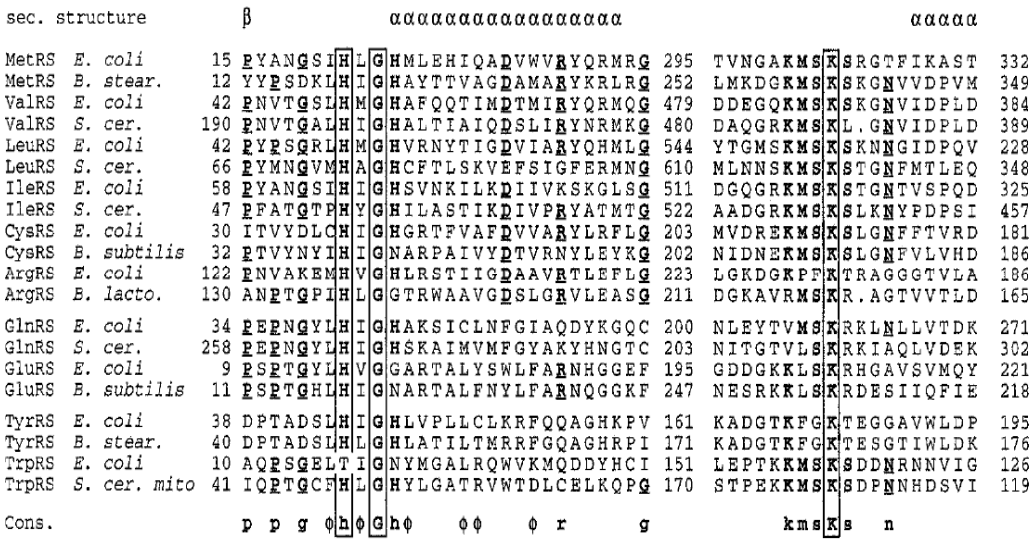


FIG. 2.5 – **Alignement des synthétases de classe I issu d'Eriani et coll. [66].** Régions comportant les deux séquences signatures. Les chiffres indiquent respectivement le début de chacune des deux séquences présentées et la fin de la séquence. La première ligne indique la structure secondaire tandis que la dernière représente la conservation des résidus. Les majuscules reflètent une conservation totale de l'acide aminé alors que les minuscules indiquent l'acide aminé majoritaire. *l* correspond aux acides aminés de petite taille (A,G,P,S,T), *0* aux hydrophobes, + aux acides aminés chargés positivement et - à ceux chargés négativement. Figure provenant de Eriani et coll. [63].

Trois sous classes furent obtenues à partir d'alignements de séquences des synthétases de ce groupe. Il est intéressant de noter que la répartition des différentes aaRS basée sur des homologies de séquences, est corrélée avec la classification des acides aminés qu'elles fixent à une exception près. En effet, la première sous classe (Ia) regroupe essentiellement des aaRS fixant des acides aminés hydrophobes, la seconde (Ib), des aaRS spécifiques des acides aminés caractérisés par une longue chaîne latérale et la troisième (Ic), celles activant des acides aminés aromatiques. Cette répartition est illustrée par la table 2.2. Enfin, d'un point de vue fonctionnel, les aaRS de classe I fixent leur acide aminé respectif au niveau du 2'OH de l'adénosine terminale de l'ARNt et se lient à ce dernier par l'intermédiaire du petit sillon [50], [175].

2.2.2 La classe II

La classe II renferme des dimères et des tétramères. Contrairement à la classe I, la structure des aaRS de classe II ne possède pas de domaine de Rossman. Cependant, elles présentent toutes une topologie similaire construite autour d'un feuillet β de sept brins antiparallèles et de quatre hélices α, l'ensemble constituant le module actif de la protéine (figure 2.4), [40], [51]. D'autre part, ces aaRS sont caractérisées par la présence de 3 motifs conservés dans le site catalytique [65], [39]. Le motif 1 est impliqué dans la dimérisation de la protéine. Il est constitué d'une longue hélice α reliée à un brin β et possède une proline invariante localisée sur le brin β. La substitution de cette dernière affecte l'activité catalytique ainsi que la dimérisation des monomères [64]. Les motifs 2 et 3 quant à eux, appartiennent au site actif. Le second motif comprend deux brins β antiparallèles connectés par une longue boucle, tandis que le troisième motif consiste en un brin β prolongé par une hélice α. L'alignement de ces trois motifs issus de différentes aaRS de classe II est présenté par la figure 2.6.

Tout comme la classe I, la classe II peut être subdivisée en 3 groupes. La première sous-classe (IIa) renferme des aaRS présentant des homologies de séquence dans le site actif et exceptée la SerRS, dans le domaine C-terminal, région impliquée dans la reconnaissance de l'ARNt. Cette sous-classe charge des acides aminés de petite taille et polaires. Les aaRS de la seconde sous-classe (IIb) se distinguent par la présence d'un module N-terminal organisé en tonneau β impliqué dans la fixation de l'ARNt. La dernière sous-classe (IIc) regroupe toutes les aaRS de classe II dont la structure oligomérique n'est pas conservée. Sur le plan fonctionnel, les aaRS de classe II, exceptées la PheRS, se distinguent de celles de classe I par le fait de fixer l'acide aminé activé au niveau du 3'OH de l'ARNt et d'aborder ce dernier au niveau du grand sillon [50], [175].

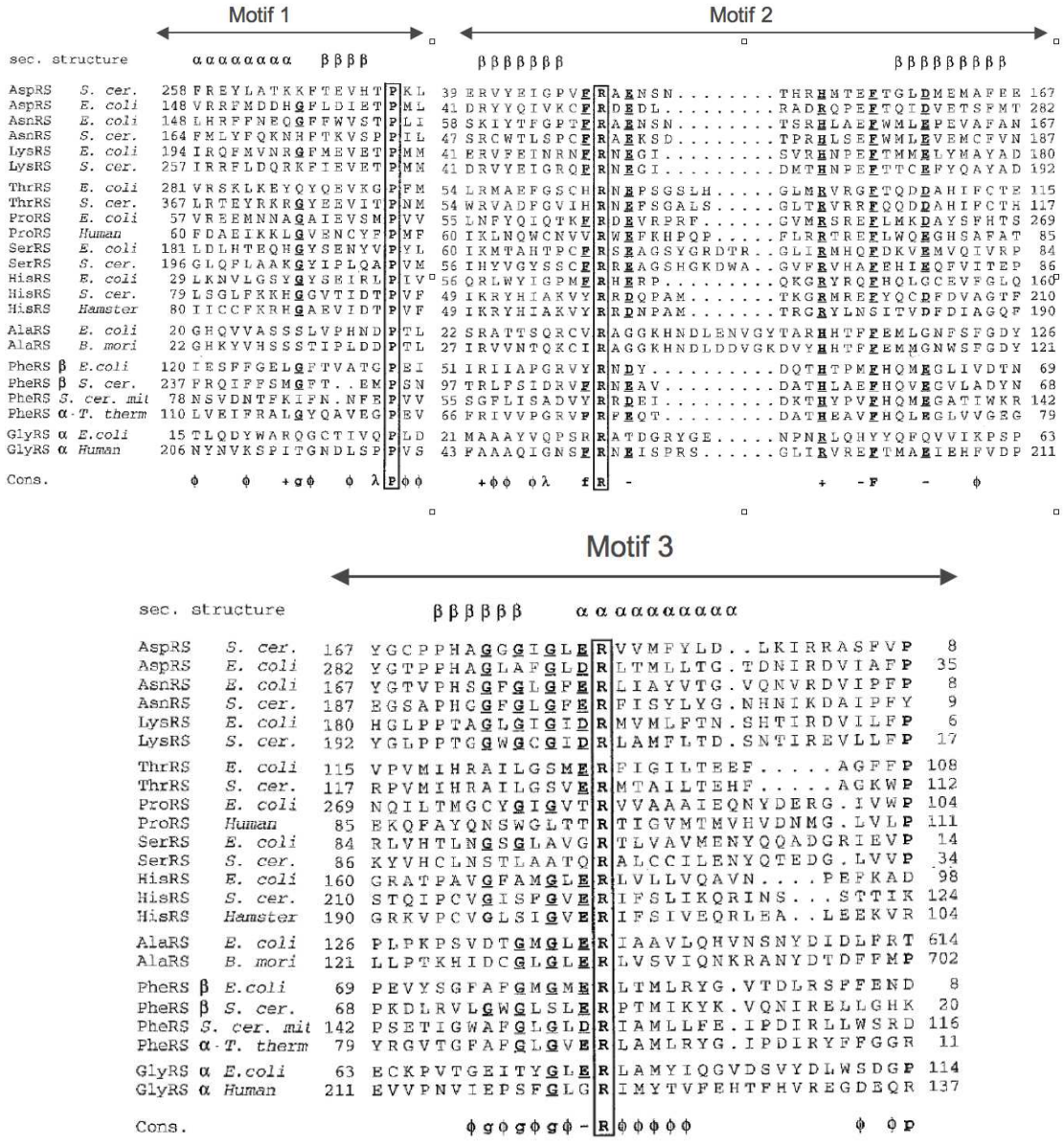


FIG. 2.6 – **Alignement des synthétases de classe II issu d'Eriani et coll. [65].** Représentation des trois motifs caractéristiques des aaRS de classe II. Les chiffres indiquent respectivement le début de chacune des séquences présentées et la fin de la séquence. La première ligne indique la structure secondaire tandis que la dernière représente la conservation des résidus. Les majuscules reflètent une conservation totale de l'acide aminé alors que les minuscules indiquent l'acide aminé majoritaire. *l* correspond aux acides aminés de petite taille (A,G,P,S,T), *0* aux hydrophobes, + aux acides aminés chargés positivement et - à ceux chargés négativement. Figure provenant d'Eriani et coll. [63].

2.3 Evolution des aaRS

Les aaRS sont des enzymes très anciens. Actuellement, on pense que leur évolution est étroitement liée avec celle du code génétique. C'est pourquoi aujourd'hui, les aaRS occupent une place prédominante dans les études réalisées sur l'origine de la vie [182], [213], [50], [49].

La théorie du monde ARN postule que les fonctions chimiques aujourd'hui accomplies par les protéines et les molécules d'ADN étaient au départ réalisées par des molécules d'ARN [76]. On pense que ces dernières étaient alors capables d'assurer l'existence de cellules primitives [173], [182]. L'hypothèse du monde ARN fut confortée par la découverte des ribozymes (pour acide RIBOnucléiques et enZYMES), molécules d'ARN capables de catalyser une réaction chimique, propriété, jusque là, exclusivement attribuée aux protéines [117]. Actuellement, cette théorie du monde ARN repose sur trois observations :

- Les molécules d'ARN et les ribonucléoprotéines jouent un rôle central dans le décodage de l'information génétique [37].

- Les molécules d'ARN peuvent jouer un rôle catalytique dans certaines réactions chimiques impliquées dans les voies métaboliques de cellules contemporaines [83]. Ainsi, Guerrier-Takada et coll. ont montré que la sous unité catalytique de la ribonucléase P correspondait au domaine ARN de l'enzyme et que cette dernière possédait une activité nucléase.

- De nombreuses équipes ont révélé une activité catalytique *in vitro* de molécules d'ARN dans certaines réactions biochimiques [62], [9].

Ces observations furent corroborées par la découverte de molécules d'ARN présentant des fonctions très proches des aaRS actuelles [95]. Par ailleurs, Lee et coll. mirent en évidence une activité catalytique chez un ribozyme synthétique alors capable d'activer un acide aminé et de l'aminoacyler à son ARNt spécifique. Ce résultat suggére un rôle possible des ARNt dans l'aminoacylation d'acides aminés [132].

Cependant, on pense que les polynucléotides furent rapidement restreints dans leurs fonctions par leur manque de diversité. Les molécules d'ARN ne peuvent pas réaliser autant d'interactions hydrophobes ou électrostatiques que les polypeptides. Cet ensemble plus restreint d'interactions disponibles à l'ARN a alors probablement promu l'émergence des protéines pour la catalyse des réactions chimiques. Le monde protéique aurait ainsi succédé au monde ARN, reléguant l'ARN à des métabolismes plus limités.

Une théorie fut proposée pour décrire le passage du monde ARN au monde protéique par l'équipe de Schimmel [173], [183], [174], [175]. Cette hypothèse attribue aux aaRS, une position déterminante dans l'émergence des protéines et les place au coeur de la transition des ARN ribosomiques au monde actuel où les acteurs de la catalyse sont essentiellement de nature protéique. Le rôle central des synthétases dans le décodage de l'information génétique expliquerait leur statut universel et laisse supposer que leur apparition fut concomitante à la mise en place du code génétique actuel [19], [159]. Les auteurs supposent que les ARNt primordiaux étaient constitués d'une unique tige-boucle acceptrice de l'acide aminé. Ces ARNt devaient dans un premier temps être aminoacylés par des ribozymes qui auraient ensuite été assistés par des protéines puis finalement remplacés par ces dernières.

Le groupe de Schimmel réalisa une étude très intéressante sur la structure des complexes aaRS-ARNt proposant alors une explication sur l'origine de la spécificité des différentes aaRS, en particulier leur découpage en deux classes [175]. L'analyse structurale des complexes synthétases-ARNt révéla que les synthétases de classe I se liaient à l'ARNt par l'intermédiaire du petit sillon tandis que celle de classe II l'abordaient par le grand sillon (figure 2.7).

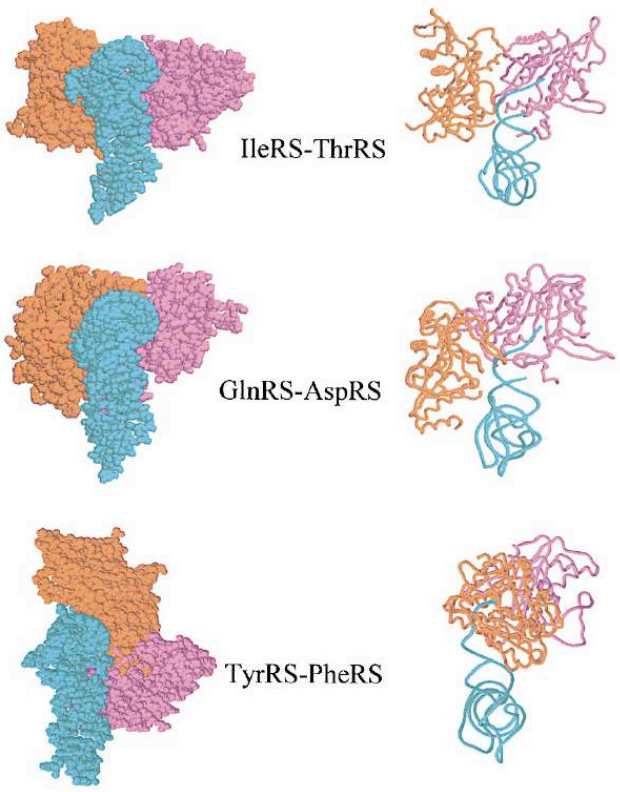
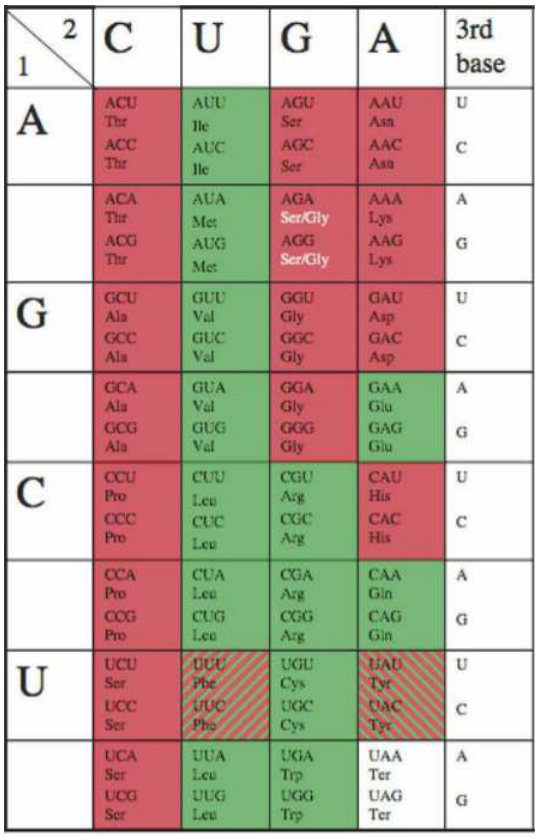


FIG. 2.7 – **Représentation de paires de synthétases compatibles de classe I et II complexées simultanément avec l'ARNt.** L'ARNt est représenté en bleu tandis que les aaRS de classe I et II sont respectivement représentées en orange et rose. A gauche : représentation 'spacefill'. A droite : représentation de la chaîne principale. Il est à noter que le complexe IleRS-ThrRS est équivalent aux complexes IleRS-SerRS, ValRS-SerRS et ValRS-ThrRS (non montrés). Figure extraite de l'article de Ribas de Pouplana et Schimmel [175].

Cette approche symétrique de l'ARNt permettrait ainsi à ce dernier de lier simultanément deux synthétases de classe opposée sans aucune gène stérique. Ces différentes orientations expliqueraient pourquoi les enzymes de classe I aminoacylerait le 2'OH de l'ARNt tandis que ceux de classe II attacherait l'acide aminé en 3'OH. Cependant toutes les paires classe I-classe II ne sont pas autorisées, certaines d'entre elles conduisant à des conflits stériques. Les auteurs montrent que les combinaisons optimales ont lieu entre des synthétases de sous classe correspondantes (Ia-IIa, Ib-IIb, Ic-IIc). Ainsi, la remarquable symétrie entre les deux classes de synthétases (dix enzymes exactement pour chaque classe, trois sous classes dans chaque groupe) pourrait découler de l'interaction de paires spécifiques de synthétases complexées simultanément à l'ARNt. Ces paires auraient alors co-évolué dans le but de recouvrir et protéger la tige-boucle acceptrice dans un environnement hostile à l'ARN (température très élevée par exemple) pour finalement conduire à deux groupes de synthétases. La réaction d'aminoacylation était alors catalysée par d'autres molécules telles que les ribozymes qui auraient été remplacés par les synthétases ensuite. Chaque membre des différentes paires d'aaRS auraient donc acquis une spécificité pour un acide aminé contribuant ainsi à la mise en place du code génétique actuel.

Delarue proposa un modèle sur le mécanisme conduisant à la mise en place du code génétique (figure 2.8) [50]. Ce mécanisme impliquerait une succession de décisions binaires permettant de réduire progressivement l'ambiguïté de tous les codons. Chaque étape de différenciation suivrait le même schéma présenté par la figure 2.9. Une des descendances serait la copie exacte de la mère et serait conservée lors de la prochaine division. L'autre descendance aurait la possibilité de se différencier à la génération suivante pour donner naissance à deux nouvelles descendances qui seraient aminoacylées par deux mécanismes différents (mécanisme caractéristique de la Classe I ou de la Classe II). Le premier événement consisterait en la différenciation

de la seconde base en purine ou pyrimidine. Cette étape conduirait à une spécification de cette base en C/U et G/A. Puis la première base se différencierait en purine ou pyrimidine, suivie d'une nouvelle étape de spécification en C/U et G/A. Enfin, la troisième base se différencierait en purine ou pyrimidine.



1 \ 2	C	U	G	A	3rd base
A	ACU Thr	AUU Ile	AGU Ser	AAU Asn	U
	ACC Thr	AUC Ile	AGC Ser	AAC Asn	C
	ACA Thr	AUA Met	AGA Ser/Gly	AAA Lys	A
	ACG Thr	AUG Met	AGG Ser/Gly	AAG Lys	G
G	GCU Ala	GUU Val	GGU Gly	GAU Asp	U
	GCC Ala	GUC Val	GGC Gly	GAC Asp	C
	GCA Ala	GUA Val	GGA Gly	GAA Glu	A
	GCG Ala	GUG Val	GGG Gly	GAG Glu	G
C	CCU Pro	CUU Leu	CGU Arg	CAU His	U
	CCC Pro	CUC Leu	CGC Arg	CAC His	C
	CCA Pro	CUA Leu	CGA Arg	CAA Gln	A
	CCG Pro	CUG Leu	CGG Arg	CAG Gln	G
U	UCU Ser	UUU Phe	UGU Cys	UAU Tyr	U
	UCC Ser	UUC Phe	UGC Cys	UAC Tyr	C
	UCA Ser	UUA Leu	UGA Trp	UAA Ter	A
	UCG Ser	UUG Leu	UGG Trp	UAG Ter	G

FIG. 2.8 – **Représentation du code génétique.** Chaque codon est coloré en fonction du mécanisme d'aminoacylation de son ARNt correspondant. Par soucis de concision, nous parlerons de codons aminoacylés/chargés en 2'OH (vert) ou 3'OH (rouge). La PheRS et la TyrRS constituent deux cas ambigus. La PheRS fait partie de la classe II tandis qu'elle est aminoacylée en 2'OH comme les autres synthétases de classe I. A l'inverse, la TyrRS appartenant à la classe I est aminoacylée en 3'OH comme les autres synthétases de classe II. Par conséquent les codons Phe et Tyr sont colorés en rouge et vert (hachage). Les codons AGR (R correspond aux purines) ont été assignés à Gly/Ser comme dans de nombreux variants de mitochondries (imprimés en blanc). Cette figure est extraite de l'article de Delarue [50].

Finalement, on pense que l'ARNt assurait seul au départ l'aminoacylation mais se trouvait limité, par son manque de diversité, à quelques acides aminés. L'entrée en jeu du monde protéique et en particulier des synthétases introduisit alors de la diversité et de la spécificité qui permirent l'aminoacylation d'un répertoire plus large d'acides aminés et par conséquent conduisirent à l'extension du code génétique.

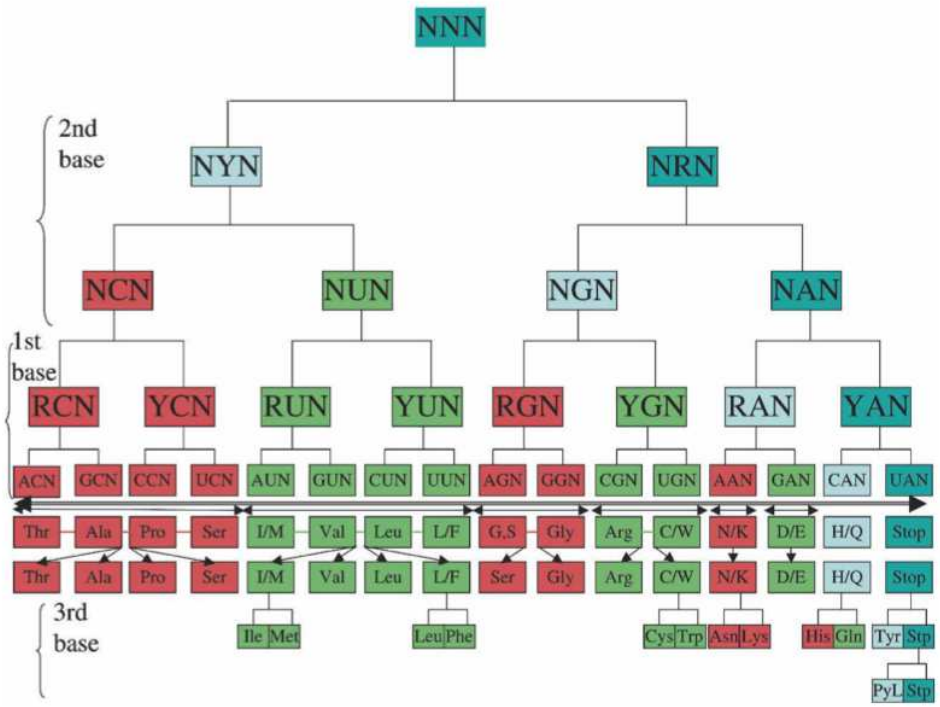


FIG. 2.9 – **Différenciations successives conduisant au code génétique actuel (figure 2.8).** A chaque évènement de différenciation, une descendance est une copie exacte du phénotype original (bleu foncé) tandis que l'autre (bleu clair) se différenciera de façon irréversible en deux codons de couleurs différentes (rouge et vert). Les purines sont représentées par un R et les pyrimidines par un Y. Ce mécanisme implique que le codon bleu foncé corresponde au codon STOP (ne peut être aminoacylé ni en 2'OH ni en 3'OH) tandis que le codon bleu clair correspond à un codon ambiguë pouvant être aminoacylé en 2'OH ou en 3'OH. Cette figure est extraite de l'article de Delarue [50].

2.4 Evolution dirigée des aaRS

Les synthétases furent l'objet de nombreuses études, d'une part pour comprendre leur rôle et leur mode de fonctionnement dans la cellule, d'autre part, pour réaliser leur ingénierie. Le principe consiste à intervenir au niveau de leur fonction d'activation ou d'édition afin de pouvoir modifier la synthèse protéique.

2.4.1 Changement de spécificité des aaRS

En 1998, le groupe de Mirande modifia pour la première fois la spécificité d'une synthétase [2]. Les auteurs s'intéressèrent aux GlnRS et GluRS, deux synthétases structuralement proches. On suppose que leur gènes proviennent d'une duplication d'un gène ancestral [122]. Bien que les mécanismes de reconnaissance de l'acide aminé des deux synthétases diffèrent, elles possèdent une organisation structurale de leur site catalytique conservée, représentative de leur ancêtre commun.

L'idée du groupe de Mirande était de créer un mutant de la GlnRS capable d'aminoacyler l'ARNt avec un glutamate au lieu d'une glutamine. L'analyse phylogénétique des sites actifs des deux enzymes suivie de leur analyse structurale révéla un groupe de résidus conservés dans la GlnRS et remplacés systématiquement dans la GLuRS. Ces résidus furent donc choisis comme cible pour modifier la spécificité de cette synthétase ; Les auteurs introduisirent alors une double mutation conduisant à un permutation notable de la spécificité de la GlnRS. Le double mutant catalysait préférentiellement l'activation de l'acide aminé Glu.

2.4.2 Extension du code génétique

Ces dernières années, un grand nombre d'études dédiées à l'extension du code génétique furent réalisées, [208], [214], [194], [38]. Encoder génétiquement des acides aminés non-naturels fournit une méthode très puissante pour l'exploration et la

manipulation de la structure et la fonction d'une protéine. Si les méthodes de mutagénèse dirigée classiques ont considérablement amélioré nos capacités à manipuler la structure des protéines, le nombre de substitutions que permet le code génétique actuel demeure limité. Cette nouvelle approche permet pour la première fois d'encoder génétiquement de nouveaux acides aminés avec des nouvelles propriétés physiques, chimiques et biologiques dans des systèmes procaryotes et eucaryotes.

Méthodologie générale

Incorporer *in vivo* un nouvel acide aminé dans le répertoire du code génétique requiert de nouveaux composants de la machinerie de la biosynthèse protéique. Il faut un codon spécifique de l'acide aminé non-naturel ainsi qu'une paire ARNt-aaRS capable, en réponse à ce codon, d'incorporer l'acide aminé non-naturel dans les protéines en cours de synthèse. Enfin, le milieu intra-cellulaire doit contenir un niveau significatif de l'acide aminé non-naturel.

La première étape consiste à choisir un codon ne codant pour aucun des 20 acides aminés classiques qui sera alors spécifique du nouvel acide aminé non-naturel. Pour ce faire, il est possible d'utiliser l'un des trois condons STOP, un quadruplet de bases d'ARN ou encore une bactérie modifiée pour laquelle les codons redondants ainsi que leur ARNt correspondants ont été éliminés du génome. On rappelle qu'un acide aminé peut être codé par plusieurs codons. Généralement, on choisit le codon STOP amber (UAG). En effet, il est le moins utilisé, parmi les trois codons STOP existants chez *E. coli* et les levures et n'est impliqué que très rarement dans des gènes essentiels. Par conséquent, son utilisation détournée ne doit pas perturber de manière significative le bon fonctionnement de l'organisme hôte.

L'étape suivante consiste à construire un nouvel ARNt 'orthogonal'. Il s'agit d'un ARNt non reconnu par les aaRS endogènes de l'organisme hôte mais capable de fonctionner correctement lors de l'étape de traduction. En outre, cet ARNt doit

répondre uniquement au codon choisi précédemment. De même, cette méthode requiert une nouvelle aaRS, orthogonale, capable d'aminoacyler l'ARNt orthogonal uniquement. Par ailleurs, cette aaRS doit être spécifique de l'acide aminé non-naturel et ne doit reconnaître aucun des 20 acides aminés classiques. L'acide aminé non-naturel ne doit être le substrat d'aucune autre synthétase endogène.

Une paire orthogonale ARNt-aaRS peut être obtenue en important une paire d'un organisme différent de l'hôte. Wang et coll. créèrent la première paire orthogonale ARNt-aaRS, en important une TyrRS et un ARNt(tyr) d'une archaebactérie, *Methanococcus Jannaschii* (*M. Jannaschii*) chez *Escherichia Coli* (*E. coli*) [207]. Les archaebactéries possèdent des composants de la machinerie de traduction assez différents de ceux d'*E. coli*. Par conséquent, ces derniers ne peuvent pas directement interagir avec les synthétases et ARNt endogènes de chez *E. coli*. Un des avantages de la TyrRS de *M. Jannaschii* est qu'elle ne possède pas de fonction d'édition permettant normalement de déacyler un acide aminé non-naturel.

Ensuite, une étape d'évolution dirigée est nécessaire pour augmenter la spécificité de la synthétase orthogonale pour l'acide aminé non-naturel. A cet effet, des librairies de variants de synthétases sont générées. Les mutations sont réalisées de façon aléatoire. Les différents mutants subissent alors une succession de sélections positives et négatives de manière à isoler des mutants capables d'incorporer spécifiquement l'acide aminé non-naturel en réponse au codon choisi au préalable.

Enfin, le nouvel acide aminé doit pouvoir être transporté du milieu de culture au cytoplasme de manière efficace. Une étude expérimentale réalisée sur des levures et *E. coli* a montré que la plupart des acides aminés non-naturels pouvaient être acheminés jusqu'à leur cytoplasme par les transporteurs d'acides aminés classiques avec la même efficacité [147].

Exemple d'application

En 2001, le groupe de Schultz, présenta pour la première fois l'ingénierie d'une paire orthogonale ARNt-aaRS capable d'incorporer *in vivo* et spécifiquement un acide aminé non-naturel dans les protéines d'*E. coli* en réponse au codon amber [209]. Ainsi, la O-méthyl-tyrosine, acide aminé synthétique, fut ajoutée avec une très bonne efficacité au code génétique d'*E. coli* [207]. D'autres études similaires furent appliquées aux levures ou encore aux mammifères [32], [179], [90].

Autre méthode

Marlière et Schimmel envisagèrent le problème par une méthode différente [59]. Les auteurs s'appuyèrent sur l'activité d'édition des synthétases. Il arrive que le site actif peine à discriminer les acides aminés très proches d'un point de vue structural. C'est pourquoi une fonction d'édition fut développée par les synthétases afin de corriger les erreurs d'amioacylation [67].

Les auteurs s'intéressèrent au cas de la ValRS connue pour activer par erreur la thréonine ou l'aminobutyrate [96]. Cette erreur peut être corrigée par la fonction d'édition de la ValRS qui conduit à l'hydrolyse de la liaison entre l'ARNt^{val} et le mauvais acide aminé. S'appuyant sur la similarité structurale de la cystéine et de l'aminobutyrate, les auteurs sélectionnèrent des mutants du site d'édition de ValRS capables de charger la cystéine sur l'ARNt^{val}. Ces mutants étaient également capables de charger l'aminobutyrate. Ainsi, plus de 20% des valines des différentes protéines cellulaires, purent être remplacées *in vivo* par l'aminobutyrate. Le fait que toutes les mutations obtenues soient localisées dans le site d'édition de la protéine souligne l'importance de cette fonction dans la restriction du code génétique à 20 acides aminés. Ceci peut alors expliquer pourquoi le site d'édition d'une synthétase donnée reste rigoureusement conservé au sein d'organismes très éloignés. Cette étude montre que modifier cette fonction d'édition peut permettre l'incorporation efficace de nouveaux acides aminés dans un organisme.

Actuellement, plus de trente acides aminés non-naturels ont été incorporés *in vivo* dans des protéines [214], [208]. Ces acides aminés peuvent conférer une nouvelle diversité aux protéines grâce à leurs nouvelles propriétés spectroscopiques, chimiques ou structurales. Cependant, ces méthodes expérimentales demeurent coûteuses, très longues à réaliser et ne permettent pas une exploration exhaustive de l'espace des séquences disponibles pour la région à muter.

L'alternative serait d'utiliser les méthodes de CPD, ce qui permettrait l'exploration plus exhaustive et contrôlée des mutations possibles plutôt que de générer expérimentalement et aléatoirement des librairies de mutants. Ainsi, nous nous sommes attelés à l'élaboration d'une procédure automatique de CPD afin de réaliser l'ingénierie d'une synthétase dont la spécificité serait modifiée. Nous nous sommes en particulier intéressés à l'AspRS et l'AsnRS, synthétases de classe II présentant de grandes similarités structurales et fonctionnelles.

2.5 Sujet d'étude : AspRS et AsnRS

Nos travaux ont porté sur l'AspRS et l'AsnRS, synthétases de la classe IIb, pour lesquelles de nombreuses études structurales et biochimiques furent réalisées [26], [169], [52], [40]. Les caractéristiques structurales de l'AspRS seule ou en présence d'aspartate, d'ATP ou d'adénylate furent largement étudiées *via* des méthodes cristallographiques par le groupe de Moras, [27], [28], [29], [26], [52], [169]. Actuellement, la Protein Data Bank recense une dizaine de structures d'AspRS.

Par ailleurs, la structure de l'AsnRS d'*E. coli* fut déterminée par cristallographie par Nakatsu et coll., tandis que, Berthet-Colominas et coll. présentèrent celle de *Thermus thermophilus* (*T.thermophilus*) [160], [161], [12]. En dépit de leur faible taux d'identité, la structure de l'AsnRS est remarquablement similaire au domaine catalytique de l'AspRS (figure 2.10).

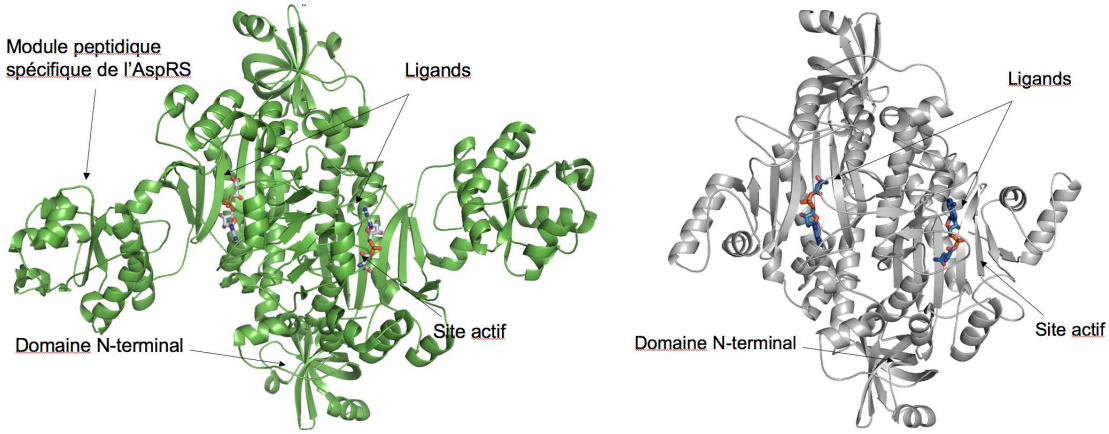


FIG. 2.10 – **Représentation en mode 'cartoon' des dimères d'AspRS (gauche) et d'AsnRS (droite).** Le ligand est représenté en mode 'stick'.

L'organisation structurale de la poche de liaison et la plupart des résidus catalytiques sont conservés au sein des deux synthétases suggérant un processus d'évolution divergente à partir d'un ancêtre commun. Nous allons dans un premier temps, présenter les caractéristiques structurales du site actif de l'AspRS puis nous porterons notre attention sur celles de l'AsnRS.

2.5.1 L'Aspartyl-ARNt synthétase

L'AspRS est un dimère composé de deux monomères d'environ 550 résidus, chacun des monomères étant constitué de deux domaines. Le domaine N-terminale structuré en baril β correspond au domaine de liaison à l'anticodon. Constitué d'une centaine d'acides aminés, il est connecté par une boucle d'une vingtaine de résidus à la région C-terminale, domaine catalytique de l'enzyme. L'interface du dimère implique les domaines C-terminaux de chaque sous unité.

Structure du site actif

Le site actif des synthétases de classe II s'avère fortement conservé au sein de ce groupe et demeure relativement rigide durant toutes les étapes de l'aminoacylation, à

l'inverse des enzymes de classe I qui présentent une poche catalytique beaucoup plus flexible [26]. Cavarelli et coll. se sont particulièrement intéressés au site d'activation de l'AspRS de la levure *Saccharomyces Cerevisae* (*S. cerevisae*) [26]. Le domaine catalytique comprend un feuillet β constitué de six brins anti-parallèles complétés par un septième brin parallèle. L'ensemble forme le 'plancher' du site actif dont l'entrée est obstruée par trois boucles peptidiques. L'aspartate se fixe sur la surface aménagée par le feuillet β (figure 2.11). Lorsque l'ATP est présent dans la poche de liaison, il empêche l'entrée de l'aspartate dans le site actif suggérant un mécanisme de réaction ordonné où l'acide aminé pénétrerait dans le site actif en premier suivi de l'ATP.

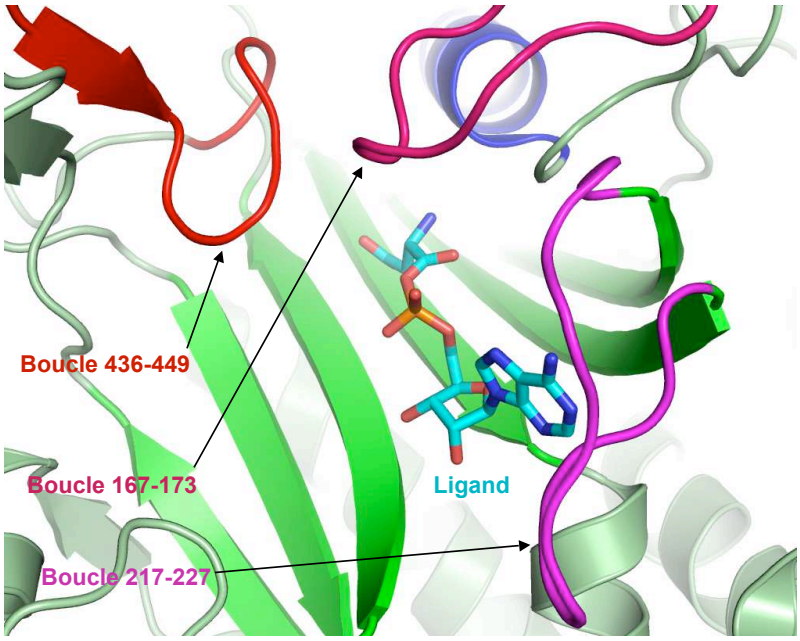


FIG. 2.11 – **Le complexe AspAMP fixé sur le plancher formé par le feuillet β.** Les trois boucles peptidiques obstruant l'entrée du site actif sont représentées respectivement en rouge, en rose et en magenta. Le ligand est représenté en turquoise. La numérotation des boucles correspond à celle de l'AspRS d'*E. coli*.

Fixation de l'ATP

Chez toutes les synthétases de classe II, l'ATP se fixe dans une conformation coudée et interagit par l'intermédiaire de ses trois phosphates avec trois ions Mg^{2+} (deux dans quelques rares cas). Le cation Mg-1, nécessaire à la première étape d'aminoacylation, vient stabiliser les phosphates α et β de l'ATP. Cet ion est lui même neutralisé par les résidus Asp 471 [475][1] et Glu 478 [482]. Les deux autres cations se positionnent de part et d'autre des phosphates β et γ. Le rôle des trois ions Mg^{2+} n'est pas clairement déterminé, mais on suppose qu'ils participent à la stabilisation de l'état de transition. Les ions Mg-2 et Mg-3 neutralisent vraisemblablement le pyrophosphate libéré lors de l'hydrolyse de l'ATP.

Par ailleurs, de nombreuses interactions conservées dans la plupart des synthétases de classes II vont stabiliser l'ATP dans une conformation coudée caractéristique des synthétases de classe II. Chez l'AspRS de levure, la base adénine de l'ATP interagit avec le cycle de la Phe 338 [229] d'un côté et avec l'Arg 531 [537] de l'autre (figure 2.12) [26]. D'autre part, la conformation coudée de l'ATP favorise la formation d'un pont salin entre son phosphate γ et l'Arg 531 [537]. L'Arg 325 [217], commune à toutes les synthétases de classe II, participe aussi à la fixation de l'ATP en interagissant avec son phosphate α. Finalement, la plupart des résidus intervenant dans la fixation de l'ATP appartiennent aux motifs 2 et 3 et engagent des interactions par l'intermédiaire de leurs chaînes latérales invariantes au sein des enzymes de la classe II (à l'exception de l'Arg 325 [217] remplacée par une histidine chez l'AlaRS d'*E. coli*).

[1] les chiffres entre crochets correspondent à la numérotation de l'AspRS d'*E. coli*

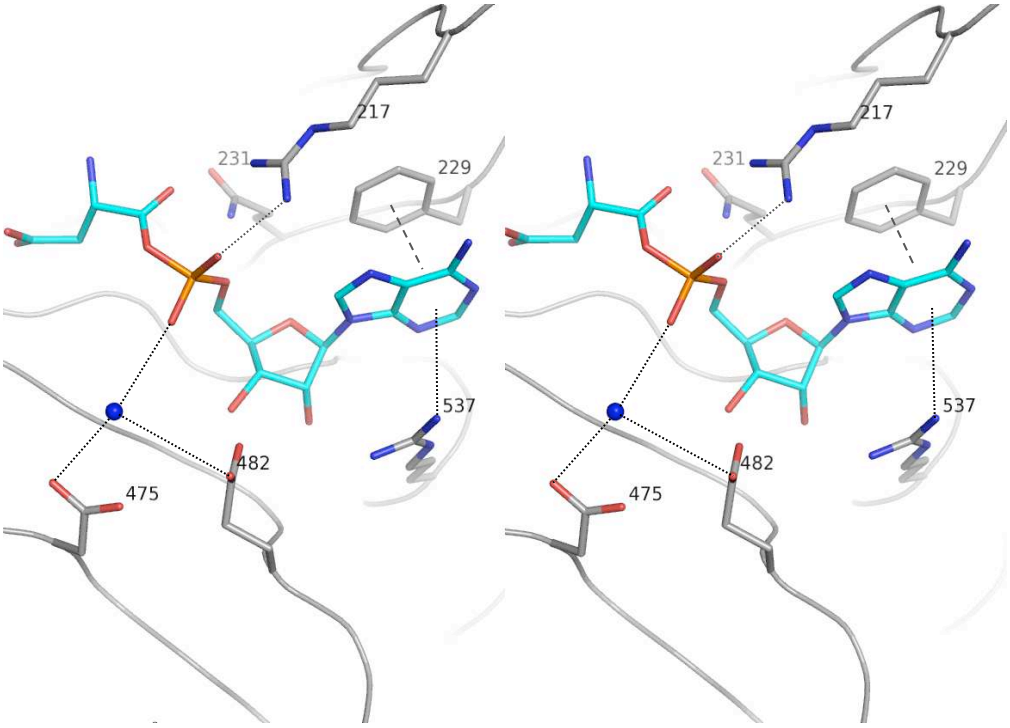


FIG. 2.12 – **Représentation stéréo de la stabilisation de la région AMP du ligand (turquoise) dans le site actif de l'AspRS d'*E. coli*.** Le cation Mg-1 qui interagit avec le phosphate α est représenté par une sphère bleue. Les liaisons hydrogènes sont matérialisées par des lignes pointillées. Le cycle benzène de la phénylalaline et le groupement guanidium de l'arginine sont parallèles et equidistants du plan de l'adénine.

Fixation de l'aspartate

L'analyse structurale de la poche catalytique montre que cette dernière est parfaitement adaptée à l'aspartate qui se lie dans la poche constituée du plancher formé par la surface du feuillet β commun à toutes les synthétases de classe II et d'un réseau de résidus conservés dans toutes les AspRS.

Chez les procaryotes, le groupement NH_3^+ engage des liaisons hydrogènes avec les résidus Glu 177 [171], Ser 199 [193] et Gln 201 [195] (figure 2.13) [169]. L'Asp 239 [233] conservé chez tous les enzymes de classe II à l'exception de l'HisRS

et l'AlaRS d'*E. coli*, interagit avec le groupe amonium du substrat par l'intermédiaire d'une molécule d'eau observée dans plusieurs structures cristallines [169]. Il est intéressant de noter que chez la levure, les résidus Glu 281 [171] et Ser 301 [193], correspondant respectivement aux Glu 177 et Ser 199 de *T. thermophilus*, n'apparaissent pas en interaction avec le groupement NH_3^+ du ligand dans la structure cristallographique. Cette dernière fut effectivement déterminée en présence d'ARNt qui entre alors en compétition avec le NH_3^+ de l'aspartate pour interagir avec le Glu 281 [171] et la Ser 301 [193] [26]. D'autre part, le groupement carboxylate du squelette du ligand est stabilisé par une liaison hydrogène avec l'Arg 325 [217] conservée chez toutes les synthétases de classe II [6].

Dans la structure de l'AspRS de levure, la chaîne latérale du substrat est reconnue par l'Arg 485 [489], la Lys 306 [198], et la chaîne principale de la Gly 526 [530] (figure 2.13). Les deux chaînes latérales chargées positivement réalisent des liaisons hydrogènes qui stabilisent fortement le groupement carboxylate du ligand. L'Arg 485 est elle même maintenue dans cette orientation favorable à la reconnaissance de l'aspartate par des interactions électrostatiques avec le Glu 344 [235]. Par ailleurs, le Glu 344 [235] et l'Asp 342 [233] stabilisent aussi, *via* des liaisons hydrogènes, la Lys 306 [198] dans la bonne orientation. Ces résultats sont concordant avec les observations faîtes chez *T. thermophilus*, où le groupement carboxylate de la chaîne latérale du ligand interagit avec l'Arg 483 [489], la Lys 204 [198], ces deux derniers résidus étant eux mêmes stabilisés par l'Asp 239 et le Glu 241. Ce réseau de quatre résidus chargés est strictement conservé au sein des AspRS et est déjà dans la bonne conformation en absence du ligand [6]. D'autre part, l'analyse structurale de l'AspRS de *T. thermophilus* révèle une interaction supplémentaire engageant l'Od1 du ligand et l'His 442 [448], acide aminé présent chez tous les procaryotes [169].

Finalement, l'aspartate engagé dans un réseau de liaisons hydrogènes et de ponts salins impliquant les résidus Gln 303 [195], Arg 485 [489], Glu 344 [235], Lys

306 [198], Gln 307 [199], Asp 342 [233], Ser 301 [193] et Gln 281 [171] (numérotation de *S. cerevisae* et [*E. coli*]) [26]. Ces résidus sont strictement invariants chez l'AspRS, à l'exception de ceux interagissant par le biais de leur chaîne principale. Ils sont eux-mêmes maintenus dans des conformations favorables à la reconnaissance de l'aspartate par d'autres résidus légèrement moins bien conservés. La table 2.2 résume les principales interactions du site actif et la correspondance entre les résidus du site actif issus de différentes AspRS et AsnRS. La figure 2.13 représente les principales interactions survenant entre l'aspartate et l'AspRS.

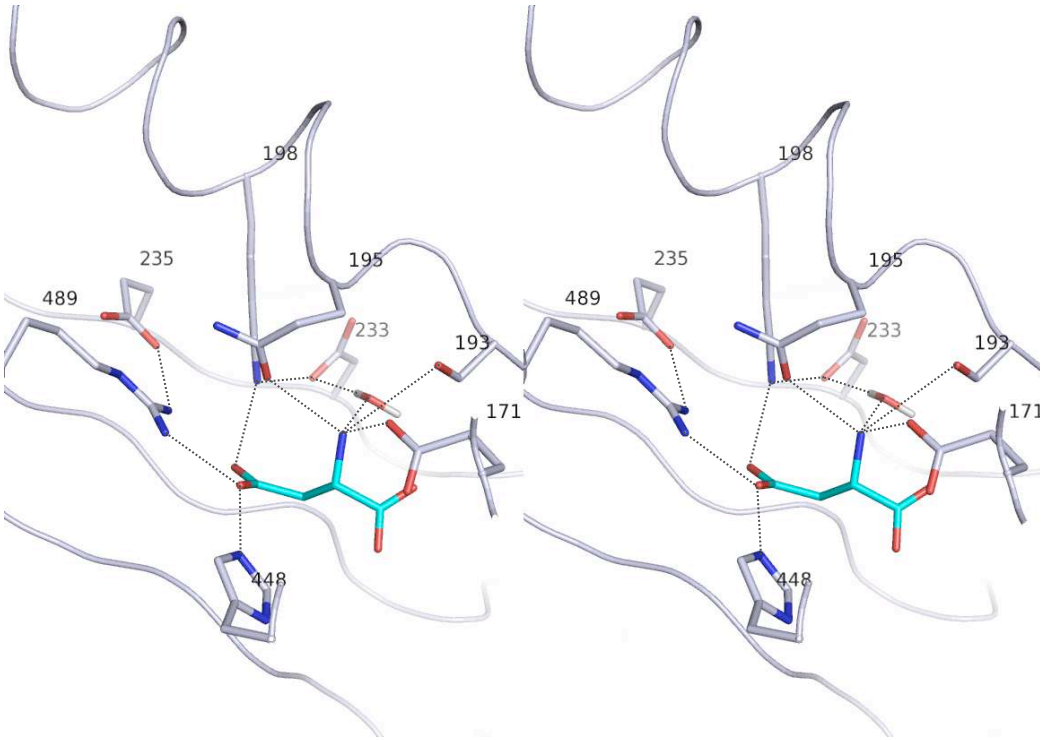


FIG. 2.13 – **Représentation stéréo des principales interactions impliquant l'aspartate dans le site actif de l'AspRS d'*E. coli*.** L'aspartate est coloré en turquoise. La molécule d'eau est représentée en mode 'sticks'. Les interactions survenant entre les différents résidus sont matérialisées par des lignes pointillées.

Res.	AspRS Sc	AspRS Ec	AspRS Tt	AsnRS Tt	Interaction avec
1	Arg 325	Arg 217	Arg 223	Arg 208	α-phosphate
2	Phe 338	Phe 229	Phe 235	Phe 221	adénosine
3	Asp 471	Asp 475	Asp 469	Asp 352	Mg-1
4	Glu 478	Glu 482	Glu 476	Glu 361	Mg-1/Mg-2
5	Glu 281	Glu 171	Glu 177	Glu 164	α-amino
6	Ser 301	Ser 193	Ser 199	Ser 185	α-amino
7	Gln 303	Gln 195	Gln 201	Gln 187	α-amino
8	Lys 306	Lys 198	Lys 204	Ala 190	chaîne latérale du ligand[a]
9	Gln 307	Gln 199	Gln 205	Glu 191	résidu (7)
10	Asp 342	Asp 233	Asp 239	Glu 225	α-amino via H20/ résidu (8)[a]
11	Glu 344	Glu 235	Glu 241	Glu 227	résidu (13)
12	-	His 448	His 442	Lys 334	chaîne latérale
13	Arg 485	Arg 489	Arg 483	Arg 368	chaîne latérale

Table 2.2 : Correspondance des principaux résidus du site actif entre les AspRS de *S. cerevisae* (Sc), *T. thermophilus* (Tt) et *E. coli* (Ec) et l'AsnRS de *T. thermophilus*. La dernière colonne indique les principales interactions du résidu de la ligne. Quand aucun numéro de résidu n'est précisé, cela signifie que le groupement indiqué dans la dernière colonne appartient au ligand. [a] exclusivement chez l'AspRS.

Validation par mutagénèse dirigée

Les conclusions tirées des analyses structurales furent confirmées par des expériences de mutagénèse dirigées menées par Cavarelli et coll.. Ces derniers mutèrent systématiquement les résidus à priori impliqués dans la liaison à l'ATP et la reconnaissance de l'aspartate afin d'évaluer l'effet de leur mutations sur la réaction d'aminoacylation chez l'AspRS de levure [26], [1]. Il ressort de cette étude, que la mutation des

résidus directement en contact avec l'ATP entraîne dans la plupart des cas, une diminution de l'activité enzymatique. Lors de la mutation de ces résidus, les mesures expérimentales révèlent une altération du Km de la réaction expliquant ainsi la perte d'affinité pour l'ATP et finalement la diminution d'activité observée. Ce résultat confirme ainsi le rôle capital des résidus Arg 325 [217] et Arg 531 [537] d'un côté et Asp 471 [475] et Glu 478 [482] de l'autre dans la fixation de l'ATP [26], [1]. En particulier, la mutation des résidus 471 et 478 affecte la réaction d'activation. Ces deux résidus garantiraient le bon positionnement de l'ion Mg^{2+} nécessaire à l'hydrolyse de l'ATP.

Ces résultats permettent ainsi d'expliquer la conservation de ces résidus au sein des synthétases de classe II, le mode de fixation de l'ATP n'étant pas spécifique d'une synthétase en particulier. D'autre part, en mutant la Phe 338 [229] ou la Gly 526 [530], les auteurs mesurent une baisse d'activité de la synthétase. Toutefois, le fait que ces deux mutations n'entraînent pas de supression totale de l'activité expliquerait le fait que ces deux résidus soient nettement moins conservés au sein de la classe II.

Par ailleurs, la mutation des résidus impliqués dans la reconnaissance de l'aspartate, (Glu 303 [195], Asp 342 [233], Arg 485 [489], Glu 344 [235] et Lys 306 [198]) conduit à une diminution importante de la réaction d'aminoacylation. Les valeurs de Km, mesurées lors de la réaction d'échange pyrophosphate, révèlent, elles aussi, une altération de la liaison à l'aspartate et des propriétés catalytiques de l'enzyme. De même, l'introduction d'une chaîne latérale de taille conséquente en position 526 généralement occupée par une glycine, conduit elle aussi, à une suppression de l'activité catalytique de l'enzyme.

Différences observées entre les procaryotes et eucaryotes

Bien qu'on observe de nombreuses conservations au sein des synthétases de classe II et en particulier parmi les AspRS, il subsiste quelques différences significatives entre celles issues d'organismes procaryotes et eucaryotes.

Tout d'abord, la Gly 526 [530] commune à tous les eucaryotes est remplacée par une alanine chez les bactéries. Cette position requiert un acide aminé de très petite taille afin de laisser de la place au substrat. Or, malgré leur taille similaire, on peut supposer que le changement Ala/Gly serait responsable de la légère différence d'orientation de l'adénylate [169]. L'autre changement majeur, consiste en la présence de l'His 442[2] interagissant avec l'Od1 de l'aspartate. Cette dernière est systématiquement substituée par une arginine chez les eucaryotes.

Changements conformationnels du site actif

Poterszman et coll. se sont particulièrement intéressés aux changements conformationnels que subissait le site actif de l'AspRS de *T. thermophilus* lors de la liaison du substrat [169]. Ils montrent alors que la reconnaissance de l'acide aminé est en partie assurée par une matrice rigide. En effet, la fixation de l'aspartate n'introduit pas ou très peu de changement conformationnel des chaînes latérales de la poche à l'exception des trois boucles mentionnées en section 2.5.1 (figure 2.14). En particulier, l'His 442 subit un changement conformationnel. Cette dernière est située sur une boucle (résidus 430-443 chez *T. thermophilus*) qui s'ouvrerait en absence de substrat afin de faciliter son entrée dans la poche et se refermerait derrière lui afin de le piéger dans le site actif. Poterszman et coll. montrent qu'après fermeture de la 'valve', l'accessibilité au solvant de l'adénylate est réduite de 41 à 16 Å^2.

Ces résultats amènent à penser que le processus de reconnaissance du ligand repose sur une cavité pré-existante, complémentaire au substrat, relevant du célèbre modèle d'interaction 'clé-serrure'.

[2]numérotation *T. thermophilus*

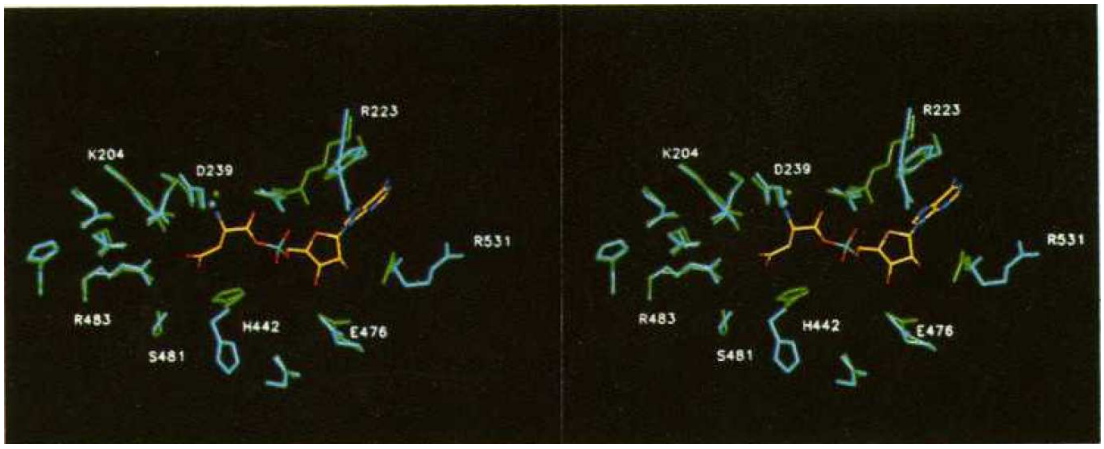


FIG. 2.14 – **Vue stéréo du site actif de l'AspRS de *T. thermophilus* en absence de substrat (bleu) et complexé avec l'AspAMP (vert).** L'AspAMP est représenté en jaune. La plupart des résidus impliqués dans l'interaction avec l'AspAMP adoptent la même conformation en absence et en présence du ligand exceptés ceux présents sur la boucle (430-443). Figure extraite de l'article de Poterszman et coll. [169].

Discrimination de l'aspartate sur l'asparagine

Bien que les séquences des sites actifs de l'AspRS et de l'AsnRS soient très similaires, l'AspRS native n'active pas l'asparagine à un niveau détectable. Archontis et coll. réalisèrent des calculs d'énergie libre sur l'AspRS en présence de l'aspartate ou de l'asparagine et calculèrent une différence d'énergie libre de liaison ($\Delta\Delta$G) de 15.3 kcal/mol en faveur de l'aspartate [6]. Différentes études furent réalisées afin de comprendre cette différence d'affinité obtenue pour deux ligands srtucturalement similaires.

La liaison du ligand requiert deux conditions : la reconnaissance précise de l'acide aminé et de bonnes énergies d'interaction entre ce dernier et la protéine. Les chaînes latérales de volume trop important sont exclues tandis que celles de petite taille telles que la glycine, alanine, valine ou thréonine sont éliminées pour ne pas réussir à réaliser des interactions assez solides avec les chaînes latérales environnantes. L'aspartate, bien que similaire à l'asparagine d'un point de vue stérique, est probablement favorisé par la présence d'une lysine en position 198 (numérotation *E.*

coli) systématiquement absente chez l'AsnRS. Cette lysine, interagit avec le groupement carboxylate de l'aspartate chargé négativement. Par ailleurs, elle est, elle même, maintenue dans une orientation favorable à la reconnaissance de l'aspartate par des interactions électrostatiques avec l'Asp 233 et le Glu 235 [26], [169]. Archontis et coll. montrèrent par des calculs PBFE (Poisson Boltzmann Free Energy) et MDFE (Molecular Dynamics Free Energy) qu'en mutant la Lys 198 en leucine, le $\Delta\Delta G_{Asp->Asn}$ devenait nul, indiquant que l'AspRS ne favorisait plus l'aspartate. La chaîne principale de la Gly 486 participe également à la stabilisation de l'aspartate, cependant, elle entre en conflit stérique avec le groupement NH_2 de l'asparagine dans le complexe AspRS :Asn [6]. La contribution individuelle de cette glycine au $\Delta\Delta G_{Asp->Asn}$ fut estimée à 6.2 kcal/mol par des calculs d'énergie libre réalisés par Archontis et coll. soulignant son rôle dans la discrimination Asp/Asn [6].

Thompson et coll. étudièrent le rôle des interactions électrostatiques dans ce phénomène de discrimination par des simulations de dynamique moléculaires menées sur l'AspRS d'*E. coli* [199], [200]. Ces interactions sont supposées jouer un rôle important dans la mesure où les trois substrats (aspartate, ATP et ARNt) sont chargés. Plusieurs analyses structurales révélent une réorganisation de la poche catalytique après la liaison de l'aspartate [188], [154]. En particulier, la boucle flexible déjà mentionnée précédement (résidus 167-173 chez *E. coli*) se referme sur le site de liaison, rapprochant ainsi le Glu 171 de l'aspartate. Ce phénomène conduit à l'introduction d'une charge négative à l'intérieur de la poche. En comparant les simulations réalisées sur les complexes AspRS :Asp ou AspRS :Asn avec une conformation ouverte ou fermée de la boucle, Thompson et coll. montrèrent que l'introduction de cette charge négative résultant de la fermeture de cette boucle défavorise l'aspartate au détriment de l'asparagine. Ce mécanisme entraîne le recrutement d'un proton labile de la part de l'His 448 qui va ainsi pouvoir interagir avec l'aspartate *via* une liaison hydrogène. Cette histidine doublement protonée va alors restaurer la discrimination en faveur de l'aspartate ($\Delta\Delta G_{Asp->Asn}$=11 kcal/mol) [199]. Les mesures expérimentales corroborèrent leur résultats supportant un état doublement protoné

pour l'His 448. Le rôle prépondérant du couplage Glu 171/His 448 dans la discrimination Asp/Asn fut confirmé par le fait qu'en maintenant artificiellement l'His 448 dans un état neutre les auteurs observèrent une baisse du $\Delta\Delta G_{Asp->Asn}$ calculé de 9 kcal/mol ($\Delta\Delta G_{Asp->Asn}$=2 kcal/mol) [199]. Par ailleurs, les auteurs répétèrent les simulations en présence du complexe ATP :Mg_3^{2+}. Le $\Delta\Delta G_{Asp->Asn}$ calculé augmenta de 11 à 19 kcal/mol suggérant un rôle important de ce complexe dans la discrimination Asp/Asn [199].

La liaison de l'aspartate à l'AspRS implique une permutation complexe entre différents états électrostatiques contrôlée par des interactions électrostatiques courtes et longues portées. Ainsi, l'ATP et le proton labile de l'His 448 agissent comme deux discriminateurs mobiles en faveur de l'aspartate, jouant un rôle essentiel dans la spécificité de l'AspRS. L'His 448 n'est pas conservée chez les eucaryotes. Néanmoins, l'ATP serait suffisant pour préserver la discrimination Asp/Asn en absence d'histidine à cette position. En effet, les auteurs calculent un $\Delta\Delta G_{Asp->Asn}$ de 9 kcal/mol lorsque l'ATP est présent et l'His 448 est maintenue dans sa forme neutre. D'autre part, la Lys 198 conservée chez toutes les AspRS, contribue également à une discrimination en faveur de l'aspartate.

Discrimination des aspartates L et D

Par ailleurs, Thompson et coll. s'intéressèrent au mécanisme mis en place par l'AspRs pour discriminer la chiralité et l'orientation de l'aspartate [198]. Les protéines sont essentiellement synthétisées à partir d'acide aminés L bien que les acides aminés D soient naturellement présents dans les cellules. Les synthétases catalysent préférentiellement l'aminoacylation des acides aminés L au détriment des D. Des mesures expérimentales révèlent que l'AspRS d'*E. coli* produit à un niveau détectable des $ARNt^{D-Asp}$, toutefois dans une quantité inférieure à celle d'$ARNt^{L-Asp}$. Les synthétases ont donc du mettre en place en mécanisme permettant de prévenir l'aminoacylation des acides aminés D.

De plus, l'aspartate, avec ses deux groupements carboxylates, présente une 'pseudo-symétrie' rompue seulement par son groupement ammonium (figure 2.15). On peut alors imaginer que l'aspartate puisse se fixer au site actif de l'AspRS dans une orientation inversée où le groupement carboxyle du squelette interagirait avec Lys 198 et Arg 489, tandis que le carboxylate de la chaîne latérale interagit avec l'Arg 217. Ainsi, l'AspRS a dû également mettre en place un mécanisme pour discriminer l'orientation native de l'aspartate de cette orientation inversée.

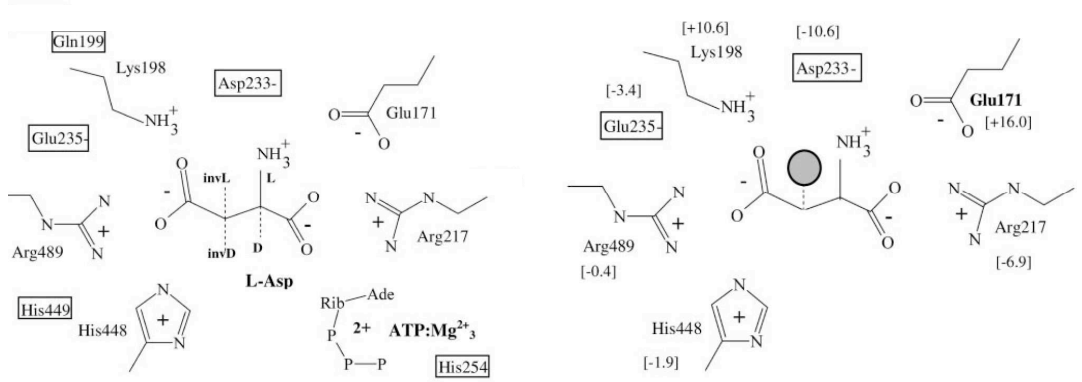


FIG. 2.15 – **A gauche :** Schéma 2D du site actif de l'AspRS d'*E. coli* en présence de l'aspartate. Les lignes pointillées indiquent les différentes positions possibles pour le groupement ammonium. L : correspond au L-aspartate dans l'orientation native, D : D-aspartate dans l'orientation native, invL : L-aspartate dans l'orientation inversée, invD : D-aspartate dans l'orientation inversée. Remarque : ces quatre ligands diffèrent uniquement dans la position du groupement NH_3^+. **A droite :** Contributions énergétiques des résidus de la poche dans le $\Delta\Delta G_{L-asp-natif->L-asp-inverse}$. L'orientation inversée du ligand correspond à un déplacement du groupement NH_3^+ du carbone α au carbone β et est matérialisée par la sphère grise. Nous constatons que lorsque l'orientation du ligand est inversée, le Glu 171 ne peut plus réaliser d'interaction stabilisante avec le groupement ammonium. Les composantes énergétiques de chaque résidus de la poche calculées avec la technique MDFE sont représentées entre crochets. Ces deux figures sont extraites de l'article de Thompson et coll. [198]

Les auteurs comparèrent L-Asp et D-Asp dans les deux orientations possibles par des calculs d'énergie libre utilisant de nouveau la technique MDFE. Leurs résultats révèlent un $\Delta\Delta G_{L-asp->D-asp}$ d'environ 6 kcal/mol. Cette valeur modérée de $\Delta\Delta G$ est en accord avec le fait qu'on observe *in vitro/in vivo* la production d'ARNt^{D-asp} en faible quantité. En revanche, le $\Delta\Delta G$ entre l'orientation native et non native du L-Asp atteint 10 kcal/mol. Ce résultat montre que l'enzyme est fortement protégé contre une orientation inversée du ligand. L'analyse des contributions individuelles des principaux résidus de la poche au $\Delta\Delta G$ observé révèle que la discrimination entre les orientations natives et inversées est principalement accomplie par Glu 171. Ce dernier, présent sur la boucle flexible (167-173) va favoriser la bonne orientation de l'aspartate en interagissant avec son groupement ammonium (figure 2.15).

Au final, la spécificité de l'AspRS est garantie par un réseau d'acide aminés en interaction conservés pour la plupart dans toutes les AspRS (exceptée l'histidine 448 absente chez les eucaryotes). Ce réseau de chaînes latérales en interaction confère à la poche, une complémentarité stérique et électrostatique avec le substrat, préservant ainsi l'AspRS des erreurs de reconnaissance du substrat. Par conséquent, la structure de la poche parfaitement adaptée à l'aspartate, permet la reconnaissance, l'orientation correcte et finalement l'activation du bon acide aminé.

2.5.2 L'Asparaginyl-ARNt synthétase

L'AsnRS est un homodimère composé de deux monomères d'environ 400 résidus. Sa structure tridimensionnelle est représentée par la figure 2.10. Comme il a déjà été mentionné précédemment, l'AsnRS présente de grandes similitudes avec les autres synthétases de classe II en particulier avec l'AspRS. Ainsi, comme chez l'AspRS, chacun des monomères est constitué de deux domaines, l'un correspondant à la

sous unité catalytique de l'enzyme, l'autre au domaine de liaison à l'ARNt. Enfin, comme chez toutes les aaRS de classe II, le domaine catalytique s'organise autour d'un feuillet β constituant le plancher du site actif sur lequel va se fixer le substrat.

Structure du site actif

Fixation de l'ATP

Berthet-Colominas et coll. présentent une analyse détaillée du mode de fixation de l'ATP qui se révèle très similaire à celui observé chez l'AspRS [12]. Comme chez l'AspRS, la superposition des structures cristallines de l'AsnRS en présence et en absence d'ATP ne révèle pas de changements structuraux majeurs, hormis une réorganisation de la boucle impliquant les résidus (210-217), phénomène déjà observé par Cavarelli [26]. Cette boucle se réordonne principalement par le biais des interactions qu'elle entreprend avec l'ATP, le stabilisant ainsi dans la même conformation courbée observée chez l'AspRS et autres synthétases de classe II [28], [10].

En particulier l'Arg 216 et l'His 217 interagissent avec le phosphate γ de l'ATP *via* des ponts salins. Par ailleurs, Berthet-Colominas et coll. décèlent la présence des trois ions Mg^{2+} retrouvés chez la plupart des aaRS de classse II. Comme chez l'AspRS, le Mg-1 crée un pont salin avec les phosphates α et β tandis que les cations Mg-2 et Mg-3 se positionnent de part et d'autre des phosphates de l'ATP pour interagir avec les phosphates β et γ. De même que chez l'AspRS, le Mg-1 est stabilisé par les résidus Asp 352 et Glu 361 correspondant respectivement aux Asp 475 et Glu 478 chez l'AspRS d'*E. coli*. L'adénine de l'ATP est elle aussi fortement stabilisée par des interactions van der Waals avec le cycle de la Phe 221 d'un côté et avec l'Arg 412 de l'autre (correspondant respectivement à l'Arg 531 et Phe 338 chez l'AspRS de levure). Les principales interactions impliquant l'ATP sont représentées par le schéma 2.16.

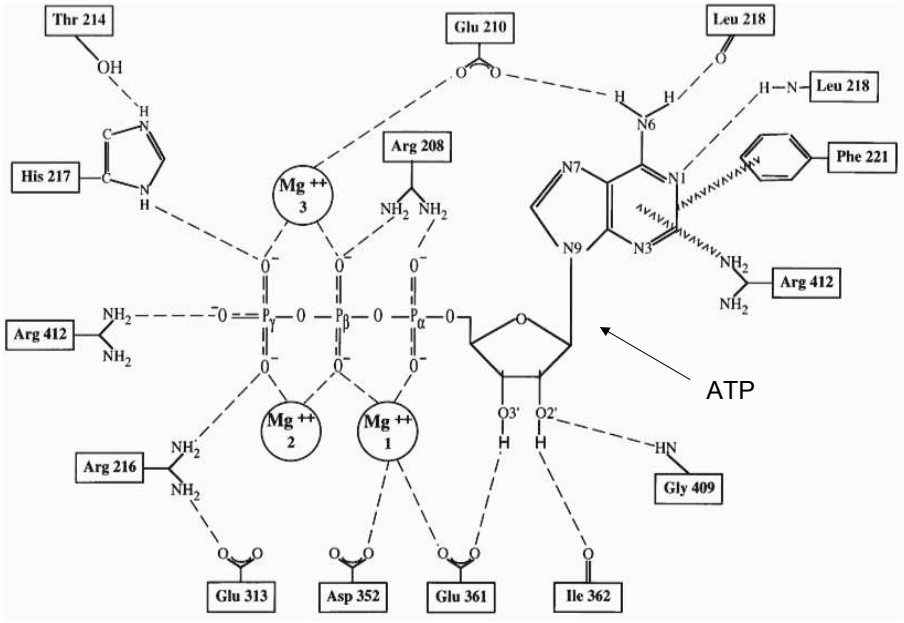


FIG. 2.16 – **Représentation 2D des principales interactions engageant l'ATP qui surviennent dans le site actif de l'AsnRS (numérotation de *T. thermophylus*).** Les liaisons hydrogènes et ponts salins sont matérialisées par des lignes pointillées, les interactions électrons Π par des lignes plus épaisses. L'ATP est représenté au centre de la figure et est indiqué par une flèche. Cette figure est extraite de Berthet-Colominas et coll. [12]

Fixation de l'asparagine

L'orientation de l'asparagine dans la poche catalytique est quasiment superposable à celle de l'aspartate chez l'AspRS. Le groupement ammonium est stabilisé par le Glu 164, la Ser 185 et la Gln 187 correspondant respectivement aux résidus Glu 171, Ser 193 et Gln 195 chez l'AspRS d'*E. coli*. Le Glu 164 fait partie de la boucle flexible qui s'avère désordonnée en absence de substrat et qui se structure ensuite de manière à obstruer l'entrée de la poche dès la fixation du ligand. L'Arg 368 (Arg 489 chez l'AspRS d'*E. coli*) conserve son rôle capital pour la reconnaissance de la chaîne latérale du ligand. La figure 2.17 représente les principales interactions survenant entre l'asparagine et l'AsnRS de *T. thermophilus*.

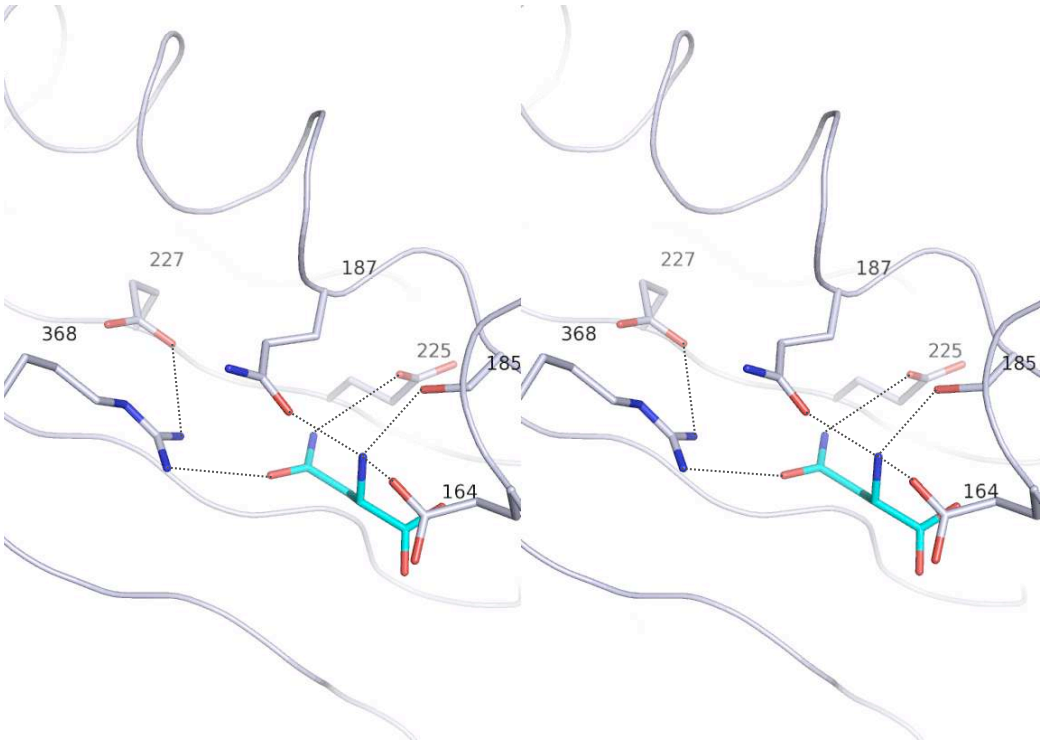


FIG. 2.17 – **Représentation stéréo des principales interactions impliquant l'asparagine dans le site actif de l'AsnRS de *T. thermophilus*.** L'asparagine est colorée en turquoise. Les interactions survenant entre les différents résidus sont matérialisées par des lignes pointillées. Le Glu 225 interagit avec le groupement NH_2 de l'asparagine, la charge négative du groupement carboxylate serait défavorable à l'aspartate.

Discrimination asparagine/aspartate

Bien que les sites actifs de l'AspRS et l'AsnRS présentent de grandes similitudes, on relève néanmoins quelques différences déterminantes dans les acides aminés du site actif contribuant sûrement à la spécificité de chaque enzyme.

Tout d'abord, la Lys 198 strictement conservée au sein de toutes les AspRS se trouve remplacée par une alanine en position 190 (numérotations correspondant

respectivement aux AspRS d'*E. coli* et AsnRS de *T. thermophilus*). Chez l'AspRS, la Lys 198, interagit avec le groupement carboxylate de l'aspartate, participant ainsi à sa reconnaissance [26]. Sa présence dans le site actif de l'AsnRS risquerait de favoriser la liaison de l'aspartate au détriment de l'asparagine. En effet, nous avons vu en section 2.5.1 que la reconnaissance de l'aspartate permet une complémentarité électrostatique entre les deux oxygènes de l'aspartate et les résidus Arg 489 et Lys 198.

Les résidus Arg 368 et Glu 227 arborent la même conformation que leur correspondants chez l'AspRS, Arg 489 et Glu 235. Cependant, l'interaction défavorable entre le groupement carboxyamide de l'asparagine et l'Arg 368 entraîne une rotation de l'axe C-Cα du substrat de 60 degrés. Cette rotation permet à l'Arg 368 d'interagir avec le groupe carbonyle de la chaîne latérale du ligand mais entraîne par ailleurs la formation d'une liaison hydrogène entre le groupement NH_2 du ligand et le résidu Glu 225. Ce dernier est conservé chez toutes les AsnRS et est systématiquement remplacé par un aspartate chez les AspRS (Asp 233 chez *E. coli*). Le Glu 225 semble jouer ainsi un rôle dans la discrimination de l'asparagine sur l'aspartate . En effet, l'aspartate se trouve désavantagé par une interaction défavorable avec le Glu 225. En revanche, dans le site actif de l'AspRS, la chaîne latérale plus courte de l'Asp 233 est plus éloignée des carboxylates de l'aspartate. Son orientation permet de faire un pont salin avec la Lys 198.

Tous ces résultats montrent qu'un réseau de charges complexe protège les sites actifs des AspRS et AsnRS contre les erreurs d'aminoacylation. La mutation systématique des résidus du site actif qui différent entre l'AspRS et l'AsnRS ne conduirait pas nécessairement à un renversement de spécificité de l'acide aminé substrat. C'est pourquoi le CPD peut jouer un rôle important dans ce domaine, en parcourant simultanément l'espace des séquences et des conformations, explorant ainsi des possibilités jusque là inenvisagées.

2.5.3 Intérêt des approches computationnelles

Les caractéristiques structurales et certains mécanismes de fixation des synthétases furent analysés avec succès par une combinaison de méthodes telles que la cristallographie, la mutagénèse dirigée, les analyses cinétiques et thermodynamiques ou encore des approches phylogénétiques. Ces méthodes se sont avérées essentielles pour nous apporter des informations sur la structure et la fonction de ces enzymes. Toutefois, au fur et à mesure que nos connaissances progressent dans ce domaine, ces approches expérimentales se trouvent limitées. En effet, elles sont restreintes aux états conformationnels les plus peuplés, les autres étant invisibles en cristallographie. Les méthodes de simulation peuvent fournir des informations complémentaires. Par exemple, les simulations de dynamiques moléculaires peuvent échantillonner des états instables structuralement ou transitoires comme le processus d'ouverture/fermeture du site actif de l'AspRS. De même, on sait que l'ATP forme majoritairement un complexe avec 3 ions Mg^{2+}, mais les données cristallographiques ne révèlent pas les occupations exactes de chacun des ions Mg^{2+}. Les simulations de dynamique moléculaire peuvent, en revanche, répondre à ce type de problème.

D'autre part, il arrive que certaines mutations s'avèrent létales ou affectent sévèrement l'activité d'une protéine. Les approches expérimentales ne révéleront pas nécessairement si l'absence d'activité provient d'une altération du mécanisme réactionnel ou tout simplement d'une altération de la liaison au substrat. En revanche, des changements d'énergie libre de liaison associés à des mutations ponctuelles peuvent être obtenus par des simulations de dynamiques moléculaires et permettre de répondre à cette question [111], [198].

Par ailleurs, les simulations permettent de décomposer l'énergie globale en contributions d'interactions microscopiques. Par exemple, dans le cas d'un complexe protéine-ligand, le $\Delta\Delta G$ associé à une mutation ponctuelle peut être décomposé en une somme de contributions de chacun des acides aminés, du ligand ou encore du solvant. Cette approche permet d'estimer le rôle de chacun des groupes dans le

processus étudié [73], [15], [198].

Ainsi, les méthodes de simulations se montrent aujourd'hui tout à fait appropriées pour des analyses structurales ou des calculs d'énergie de liaisons sur des protéines telles que les synthétases.

2.5.4 Le travail de ma thèse

Mon travail a consisté en l'élaboration d'une procédure automatique de CPD qui fut ensuite appliquée à l'évolution dirigée de l'AsnRS.

La première partie de ma thèse consista donc en un travail méthodologique. La première phase fut de mettre en place la fonction d'énergie. Il fallut répondre à trois problèmes majeurs du CPD : (1) prédiction de l'orientation des chaînes latérales, (2) solvatation, (3) mutagénèse. Dans un premier temps, nous avons implémenté le champ moyen et calibré notre fonction d'énergie de façon à prédire l'orientation des chaînes latérales. Afin d'être en mesure de comparer deux séquences pour un repliement donné, nous avons ensuite optimisé la fonction d'énergie pour la prédiction de changements de stabilité associés à des mutations ponctuelles. Lors de cette étape, trois modèles de solvant implicites furent testés et de nombreuses paramétrisations du champ de force furent comparées. Enfin, nous avons évalué notre procédure pour différentes applications incluant la prédiction d'affinité protéine-ligand et le *design* complet de protéines.

Durant la seconde période de ma thèse, j'ai appliqué notre procédure de CPD à l'évolution dirigée de l'AspRS et de l'AsnRS. Notre objectif principal était de réaliser le *design* du site actif de l'AsnRS de façon à ce qu'elle lie préférentiellement l'aspartate au détriment de son acide aminé natif, l'asparagine. Quelques ajustements de la fonction d'énergie furent nécessaires. En effet, le site actif de l'AsnRS est régi par des contraintes structurales devant aussi répondre à une nécessité fonctionnelle. La fonction d'énergie fut donc réajustée à partir du système de l'AspRS qui présente de grandes similitudes avec celui de l'AsnRS mais qui est mieux caractérisé

du point de vue expérimental. En particulier, nous avons évalué notre procédure dans sa capacité à reproduire des séquences à caractère natif ou à prédire des changements d'affinité associés à des mutations ponctuelles. Enfin, nous avons réalisé l'ingénierie du site actif de l'AsnRS. Nous avons du mettre en place une procédure pour classer les mutants obtenus. Seuls ceux présentant une stabilité raisonnable et une meilleure affinité pour le ligand non natif furent retenus. L'analyse détaillée de leurs structures 3D fut réalisée afin d'extraire une dizaine de mutants prometteurs. Leur stabilité et leurs énergies d'association avec l'aspartate sont actuellement étudiées par des simulations de dynamique moléculaire. Ces mutants seront ensuite testés expérimentalement.

Chapitre 3

Mise en place de la fonction d'énergie

3.1 Placement des chaînes latérales et mutagenèse en solvant implicite

La première partie de ma thèse consista en un travail méthodologique visant à mettre en place notre procédure automatique de CPD, en particulier, la fonction d'énergie. Nous nous sommes concentrés sur trois problèmes constituant les principaux ingrédients du CPD : prédiction de l'orientation des chaînes latérales, solvatation et mutagenèse.

3.1.1 Placement des chaînes latérales

Les algorithmes de CPD doivent sélectionner des séquences fonctionnelles parmi un grand nombre de séquences non-fonctionnelles. Ainsi, la complexité du CPD prohibe une description continue des conformations du squelette peptidique et des chaînes latérales et impose au contraire une discrétisation de l'espace conformationnel. Les chaînes latérales sont modélisées par un jeu de conformations discrètes appelées

rotamères. Les algorithmes d'optimisation vont alors déterminer dans cet espace conformationnel discret, la ou les structures les plus stables.

La première étape de ma thèse consista en l'implémentation du champ moyen afin d'optimiser notre fonction d'énergie. Nous calibrâmes notre champ de force de façon à prédire avec une bonne exactitude l'orientation des chaînes latérales d'un jeu test de 29 protéines [134]. Les calculs furent réalisés avec le modèle de solvant CASA (Coulomb Accessible Surface Area). Différentes valeurs de constantes diélectriques furent testées. D'autre part, les paramètres de solvatation atomique de Fraternali [70], Ooi [164] et Wesson [212] furent évalués. L'analyse des résultats montra que le terme de surface avait trop de poids par rapport aux autres termes de la fonction d'énergie. Différents coefficients α furent ainsi testés pour pondérer ce terme. Il ressortit de cette étude, que les paramètres de Fraternali conduisaient aux meilleurs résultats. Par ailleurs, la qualité de la prédiction des chaînes latérales est insensible aux valeurs de la constante diélectrique utilisée. Finalement, avec l'utilisation du champ moyen, un modèle de solvant très simple et des ajustements empiriques, les résultats obtenus furent de qualité comparable à d'autres travaux menés précédemment par d'autres équipes [108], [215].

3.1.2 Mutagenèse et solvatation

Dorénavant, nous étions capables de prédire avec une bonne exactitude l'orientation des chaînes latérales. Toutefois, notre algorithme de CPD doit aussi pouvoir identifier la ou les meilleures combinaisons séquences/structures. Pour tester notre capacité à comparer deux séquences, nous avons étudié l'effet de mutations ponctuelles sur la stabilité d'une protéine [134]. Le solvant joue un rôle déterminant dans la structure et la stabilité des protéines. Nous avons donc testé trois modèles de solvant implicites (CASA, GB/ACE [181] et GB/HCT [86]) pour la prédiction des changements de stabilité. Un grand nombre de paramétrisations de ces modèles furent comparés. Pour le modèle CASA, nous avons utilisé le jeu de paramètres de solvatation atomique

de Fraternali. Ce jeu très simple utilise deux paramètres uniquement. Un coefficient négatif est appliqué aux atomes polaires ou chargés tandis qu'un coefficient positif est utilisé pour les autres. Nous avons introduit un troisième paramètre pour les atomes impliqués dans les groupes ionisés. Plusieurs constantes diélectriques et différents facteurs de pondérations furent aussi comparés. Par ailleurs, le modèle GB/ACE fut légèrement modifié, tandis que le modèle GB/HCT fut complètement reparamétré pour cette application. Toutes ces optimisations furent réalisées à partir d'un jeu d'environ 1000 mutations. Comme il n'existait pas de mesures expérimentales pour ces mutations, le modèle de Poisson Boltzmann fut utilisé comme modèle de référence [134].

Ensuite, le modèle CASA et les deux modèles GB furent testés pour la prédiction de changements de stabilité associés à 140 mutations ponctuelles pour lesquelles des données expérimentales étaient disponibles. Les trois modèles de solvant donnèrent des résultats en bon accord avec les résultats expérimentaux, avec des erreurs moyennes de 2,1 kcal/mol pour CASA, 2,9 kcal/mol pour GB/ACE et 2,1 kcal/mol pour GB/HCT. Il est surprenant de voir que le modèle très simple de CASA donne des résultats de la même qualité que le meilleur des deux modèles GB. Ce résultat est encourageant pour l'utilisation du modèle CASA dans notre procédure de CPD. Ce travail est décrit dans l'article ci dessous :

A. Lopes, A. Alexandrov, C. Bathelt, G. Archontis and T. Simonson. Computational sidechain placement and protein mutagenesis with implicit solvent models, 2007. *PROTEINS : Structure, Function and Bioinformatics, 67 :853-867.*

PROTEINS: Structure, Function, and Bioinformatics 67:853–867 (2007)

Computational Sidechain Placement and Protein Mutagenesis With Implicit Solvent Models

Anne Lopes,[1†] Alexey Alexandrov,[1†] Christine Bathelt,[1] Georgios Archontis,[2*] and Thomas Simonson[1*]
[1]*Laboratoire de Biochimie (UMR CNRS 7654), Department of Biology, Ecole Polytechnique, 91128, Palaiseau, France*
[2]*Department of Physics, University of Cyprus, Nicosia, Cyprus*

ABSTRACT **Structure prediction and computational protein design should benefit from accurate solvent models. We have applied implicit solvent models to two problems that are central to this area. First, we performed sidechain placement for 29 proteins, using a solvent model that combines a screened Coulomb term with an Accessible Surface Area term (CASA model). With optimized parameters, the prediction quality is comparable with earlier work that omitted electrostatics and solvation altogether. Second, we computed the stability changes associated with point mutations involving ionized sidechains. For over 1000 mutations, including many fully or partly buried positions, we compared CASA and two generalized Born models (GB) with a more accurate model, which solves the Poisson equation of continuum electrostatics numerically. CASA predicts the correct sign and order of magnitude of the stability change for 81% of the mutations, compared to 97% with the best GB. We also considered 140 mutations for which experimental data are available. Comparing to experiment requires additional assumptions about the unfolded protein structure, protein relaxation in response to the mutations, and contributions from the hydrophobic effect. With a simple, commonly-used unfolded state model, the mean unsigned error is 2.1 kcal/mol with both CASA and the best GB. Overall, the electrostatic model is not important for sidechain placement; CASA and GB are equivalent for surface mutations, while GB is far superior for fully or partly buried positions. Thus, for problems like protein design that involve all these aspects, the most recent GB models represent an important step forward. Along with the recent discovery of efficient, pairwise implementations of GB, this will open new possibilities for the computational engineering of proteins. Proteins 2007;67:853–867.**

Key words: structure prediction; solvation; mean field; Generalized Born; Poisson equation; protein design

INTRODUCTION

Homology-based structure prediction and computational protein design are areas whose importance is increasing with the development of structural genomics.[1–4] Both techniques usually rely on a simplified description of protein conformational space, taking into account one or a few fixed backbone conformations and a discrete set of sidechain rotamers.[5–8] The most stable structure within this discrete space can be found by exact or approximate search methods.[6,9–15] This paper focusses on another important ingredient: the energy or scoring function and, in particular, the treatment of aqueous solvent. We consider the performance of several implicit solvent models for two key problems that occur in computational protein design: sidechain placement and the calculation of stability changes due to point mutations. Both problems have been extensively studied, using a range of models.[13–20] However, they must be considered together if one is to parameterize and test solvent models for protein design; i.e., for searching sequence and conformational space simultaneously. For example, previous solutions of the sidechain placement problem that omit electrostatics[14,16,21] are not acceptable in this context. Such combined analyses are much less common.[15,19,20] Furthermore, with the rapid progress of implicit solvent models, especially generalized Born models, it is important to reconsider this problem.[22,23]

Indeed, aqueous solvent plays an important role in the structure and stability of proteins.[24] Structure prediction and protein design are done almost exclusively with "implicit" solvent models, for efficiency. The solvent degrees of freedom are not explicitly represented; rather, they are taken into account through their effect on the intraprotein interactions.[25] The energy function for the protein is referred to as a Potential of Mean Force (PMF), or "effective" energy function. To obtain correct energetics for the protein, the PMF for any given protein conformation should coincide with the Boltzmann

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/fpages/0887-3585/suppmat/

[†]These authors contributed equally.

*Correspondence to: Thomas Simonson, Laboratoire de Biochimie (UMR CNRS 7654), Department of Biology, Ecole Polytechnique, 91128, Palaiseau, France. E-mail: thomas.simonson@polytechnique.fr or Georgios Archontis, Department of Physics, University of Cyprus, Nicosia, Cyprus. E-mail: archonti@ucy.ac.cy

Received 4 August 2006; Revised 26 October 2006; Accepted 16 December 2006

Published online 8 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21379

average of the energy over the solvent configurations.[25] In practice, only approximate PMFs can be constructed.

An implicit solvent model that has a clear physical basis is the Poisson model, which treats the solvent as a dielectric continuum,[26–29] by numerically solving the Poisson equation (PE). The essential physical ingredients are (1) the strong, attractive interactions between charged protein groups and the surrounding, high-dielectric solvent, and (2) the large shielding of protein–protein electrostatic interactions by solvent. The solvent contribution to the PMF is obtained as the electrostatic free energy of a collection of point charges in a dielectric cavity.[25,29] The PE model provides good accuracy for many applications,[30] including small molecule solvation,[31,32] acid/base equilibria,[28,33] ligand binding,[34] protein–protein binding,[35,36] and protein dynamics.[37] Unfortunately, PE methods cannot be used routinely for computational protein design. Indeed, in continuum electrostatics, the effective interaction between two protein residues depends on the entire protein's shape and the complementary volume occupied by high-dielectric solvent. Therefore, continuum electrostatic energies are many-body quantities that cannot ordinarily be expressed as a sum over residue or atom pairs.[19,20,23,29,38] This is a prohibitive limitation for protein design.

A more efficient alternative is the generalized Born (GB) model.[22,30,39,40] GB is based on the same physical picture as PE, with a dielectric continuum solvent surrounding a protein cavity. But it makes additional approximations that allow an analytical expression of the PMF. It has become feasible to use GB in a protein design context, because residue–pairwise implementations were recently discovered.[19,23] Several GB variants and parameterizations exist.[41–46] GB has been used for many applications, including small molecule solvation,[42–44,47] protein solvation,[45,48] acid/base equilibria,[49,50] protein dynamics,[51,52] ligand binding,[53] protein folding,[54,55] loop structure prediction,[56] and scoring native folds vs. decoys.[57] The best variants have an accuracy that is not much inferior to explicit solvent models,[22,46,49] and further improvements can be expected.

A third, even simpler class of implicit solvent models are the so-called Accessible Surface Area models.[58–60] These models characterize different atom types by "atomic solvation parameters," which reflect their hydrophobicity or hydrophilicity, and consider the fraction of each atom's surface area that is accessible to solvent. Each atom contributes to the PMF through the product of its solvation parameter and its solvent accessible surface area. Usually, this accessible surface area contribution is supplemented by another electrostatic term, which attempts to capture the shielding of protein–protein electrostatic interactions by the high-dielectric solvent. The simplest approach is to add to the PMF a screened Coulomb energy, so that protein–protein electrostatic interactions are reduced by a constant factor ε. We refer to this as the Coulomb/Accessible Surface Area (CASA) model. This class of models has been used for protein molecular dynamics,[60–62] structure prediction,[14] and protein–ligand binding.[63] It is routinely used for computational protein design.[4,64–67]

With the ongoing development of GB models, and with the discovery of methods to implement them efficiently in computational protein design,[19,23] it is important to systematically compare the behavior of the PE, GB, and CASA models for protein structure prediction and design. In the context of protein design, structure prediction is usually limited to sidechain placement with a fixed protein backbone. Many authors have analyzed this problem and shown that good results are obtained with very simple models, without any electrostatic or solvent treatment.[14,16,21] In computational protein design, however, sidechain placement must be done repeatedly, following rounds of random mutagenesis. The mutations will often modify the protein charge, frequently introducing charged sidechains into buried positions. For these effects, an accurate electrostatic treatment is desirable. Therefore, an integrated treatment of sidechain placement and ionized mutations is needed. Previous solutions of the placement problem that omit electrostatics are not acceptable. Many authors have studied the accuracy of PE and GB models for protein electrostatics.[18,29,30,68] However, very few studies have considered sidechain placement and ionized mutations simultaneously. Even fewer (none that we are aware of) have compared surface area models (CASA) to GB in this context. While CASA is the most common solvent model in computational protein design, GB models have greatly progressed in the last few years, so that a reevaluation and comparison of models is needed. For example, one recent study considered a suboptimal GB/ACE parameterization.[20]

In this paper, we consider both sidechain placement and the effect of amino acid mutations on protein stability. For sidechain placement, we use the CASA model. In a related test, we use CASA, GB, and PE to estimate the stability of large libraries of protein conformations, where the sidechain positions have been randomized. We then consider over 1000 mutations involving ionized sidechains, and use CASA, GB, and PE to compute the corresponding stability changes. The data set includes 140 mutations in 12 proteins for which experimental measurements are available. The experimental data span only a limited free energy range, and they do not include any mutations that introduce or delete ionized sidechains in the protein core, since these are usually not experimentally tractable.[69] The other mutations are mostly in buried positions and correspond to much larger stability changes, but no experimental measurements are available. For computational reasons, most of these mutations are not "biochemically exact": they consist in charge modifications, as opposed to real mutations. For example, negative charges are introduced onto valine sidechain methyls, roughly mimicking a valine → aspartate mutation. Nevertheless, they contain the relevant physical effects and represent a valid test of the solvent models. We refer to these two data sets as the "experimental" and "artificial" mutations, respectively.

An important aspect of this work is the parameterization of the CASA and GB models. With CASA, for example, we consider three sets of atomic surface parameters, a wide range of dielectric constants, and a range of scaling factors for various model terms. We consider two GB models: GB/ACE,[42] in combination with the Charmm19 force field,[70] and GB/HCT,[71] in combination with the AMBER force field.[72] A limited parameter optimization is done for GB/ACE and a more extensive one for GB/HCT. An essential point is that the parameterization must be useful for both structure prediction and stability changes. The goal of the GB/HCT parameter optimization is to take into account very recent improvements in the set of atomic radii used for continuum electrostatics calculations with the AMBER atomic charges.[73]

The "experimental" mutations are not used in the parameterization, for several reasons. First, they correspond to very small stability changes. More importantly, they cannot be computed without assuming a specific model for the unfolded protein's structure. Thus, for a rigorous optimization, the unfolded state model and the energy parameters would have to be tested and varied simultaneously. We prefer to optimize the energy function separately; therefore, we use only the artifical mutations for parameter optimization. Since experimental measurements are not available, we take as reference values the PE results. Indeed, PE has been extensively used to study protein electrostatics.[29,68] Many recent studies have shown that it is a valid reference for the parameterization and testing of implicit solvent models.[17–19] With this strategy, the unfolded state treatment does not influence the parameterization, since the same treatment is used for CASA, GB, and PE.

Once the energy parameters are chosen, we can use the experimental mutations for a blind test of the CASA and GB models. We use a very simple, tripeptide representation of the unfolded state, with noninteracting amino acids, which is commonly used in protein design. With CASA, the mean unsigned deviation from experiment is 2.1 kcal/mol. This appears comparable to the accuracy reported by Serrano et al.[74] for a comparable set of mutations, although they did not describe their subset of charged mutations separately. Remarkably, it is similar to the accuracy of molecular dynamics free energy simulations using explicit solvent models,[49,75] which are much more difficult and expensive (but which give additional structural and dynamical information). With GB/ACE and GB/HCT, the mean unsigned deviations from experiment are 2.9 and 3.4 kcal/mol, respectively. If a surface term is included in the GB/HCT model, to help represent dispersion and hydrophobic contributions,[32] the GB/HCT mean unsigned error drops to 2.1 kcal/mol, the same level as CASA.

In summary, we find that (a) the choice of electrostatic model is not very important for the sidechain placement calculations (confirming several earlier studies[14,16,21]); (b) GB/HCT and CASA give the same accuracy for the experimental mutations; (c) GB/HCT yields an enormous improvement for the total solvation free energies and for the artificial mutations. Thus, for problems like protein design that involve all these aspects, the most recent GB models represent an important step forward. Along with the recent discovery of efficient implementations of GB in protein design,[19,23] this will open new possibilities for the computational engineering of new proteins.

MATERIALS AND METHODS

Effective Energy Function

We tested several effective energy functions, or PMFs, which have the form:

$$E = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihe}} + E_{\text{impr}} + E_{\text{vdw}} + E_{\text{coul}} + E_{\text{solv}}. \quad (1)$$

The first six terms in Eq. (1) represent the protein internal energy. They are taken from either the CHARMM19 or the AMBER empirical energy function.[70,72] They represent a covalent bond energy, a bond angle energy, a torsion energy associated with sidechain dihedrals, a term maintaining the chirality or planarity of selected atomic centers, a van der Waals energy, and a Coulomb electrostatic energy,

$$E_{\text{coul}} = \sum_{i<j} \frac{q_i q_j}{r_{ij}}, \quad (2)$$

where the sum is over all pairs of protein atoms i, j, of charges q_i, q_j, and r_{ij} is the distance between a pair. The last term on the right of Eq. (1), E_{solv}, represents the contribution of solvent.[25] In this work, we compare three different solvent treatments, described below.

CASA Solvent Treatment

Our first solvent treatment is an accessible surface area treatment: the CASA model. E_{solv} includes two energy terms that describe protein–solvent electrostatic and hydrophobic interactions:[25]

$$E_{\text{solv}} \equiv E_{\text{CASA}} = E_{\text{screen}} + E_{\text{surf}} = \left(\frac{1}{\varepsilon} - 1\right) E_{\text{coul}} + \alpha \sum_i A_i \sigma_i. \quad (3)$$

E_{screen} is a screened Coulomb energy; ε is the dielectric constant of the medium (relative to vacuum). Notice that $E_{\text{coul}} + E_{\text{screen}} = E_{\text{coul}}/\varepsilon$. E_{surf} is related to the atomic solvent-accessible surface areas. The sum is over all protein atoms i; A_i is the solvent-accessible area of atom i; σ_i is an atomic solvation parameter (measured in kcal/mol/Å^2); and α is an overall weight for the surface energy term. The coefficients σ_i reflect the preference of particular atom types to be exposed or hidden from solvent, and incorporate both electrostatic and nonelectrostatic effects.[25] Surface areas were computed by the Lee and Richards algorithm,[76] implemented in XPLOR,[77] using a

1.4 Å probe radius. Three different sets of atomic solvation parameters were tested,[59–61] along with different values of the dielectric constant ε and the weight α.

GB Solvent Treatment

Our second solvent treatment is a GB model[39,41,42]:

$$E_{\text{solv}} \equiv E_{\text{GB}} = \sum_i \Delta G_i^{\text{self}} + \sum_{i<j} \Delta G_{ij}^{\text{screen}} = \tau \sum_i \frac{q_i^2}{2b_i} + \tau \sum_{i<j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j exp[-r_{ij}^2/(4b_i b_j)]}} \quad (4)$$

where $\tau = 1/\varepsilon_w - 1/\varepsilon_p$, r_{ij} is the distance between charges q_i and q_j, b_i is the effective Born radius of atom i, ε_w is the dielectric constant of water (set to 80), and ε_p is the protein dielectric constant (set to 1 unless otherwise mentioned). The first term is a sum of atomic self-energies $\Delta G_i^{\text{self}} = \tau {q_i}^2/2b_i$, corresponding to the interaction of each atomic charge q_i with its own reaction field in the environment of the solvated biomolecule. The second term models the interaction of a charge q_i with the reaction field produced by a different charge q_j, and accounts for the screening of electrostatic interactions by the high-dielectric solvent.[39]

The self-energy term in Eq. (4) requires the calculation of an electrostatic energy density integral over the solute volume. We use both the GB/ACE and the GB/HCT models, which assume a Coulomb functional form for the electric field inside the solute and partition the volume into atomic contributions.[42,71] The GB/HCT parameters are described further on. The GB/ACE atomic volumes corresponded to the Voronoi database V01,[78] scaled by a factor of 0.8 as described in Ref. 52. The smoothing parameter that controls the width of atomic volumes[42] in GB/ACE was set to 1.3. The partial charges corresponded to the CHARMM19 energy function.[70] The calculations were performed with the XPLOR program.[77,79]

Poisson Equation Solvent Treatment

Our third solvent treatment is a continuum electrostatic model with numerical solution of the Poisson equation. We refer to it as the PE model. The solvation energy has the form

$$E_{\text{solv}} \equiv E_{\text{PE}} = \frac{1}{2} \sum_i q_i V_i^{\text{reac}} \quad (5)$$

where q_i is the charge of atom i and V_i^{reac} is the electrostatic potential on atom i due to the polarization charge at the protein–solvent interface, induced by every atomic charge. For a given protein structure, we computed V_i^{reac} by solving the Poisson equation numerically for the protein in solution and in the gas phase and taking the difference. The finite-difference program UHBD was used.[80] The protein–solvent dielectric boundary was defined by the molecular surface of the protein. The solution used a two-step focussing procedure and a cubic grid with spacings of 0.8 and 0.4 Å. The molecular surface was constructed with 2000 points per atom, using a probe sphere of radius 2 Å and the boundary smoothing method in UHBD.[80] Small voids in the interior of the rotameric structures were filled by dummy atoms, to prevent the occurence of artificial, high-dielectric internal cavities (sometimes overlooked in PE applications). Two PE parameterizations are used. The first uses Charmm19 atomic radii and charges,[70] except for hydrogen radii, which were set to 1 Å. The second uses atomic charges from the AMBER, all-atom force field,[72] along with atomic radii specifically and carefully optimized for PE with AMBER charges.[73]

Protein Set and Rotamer Library

Twenty-nine proteins were used for sidechain placement calculations; they are listed in Table III. Their sizes varied from 36 to 212 amino acids, with a mean of 110. Their structures were all determined using X-ray crystallography with a resolution of 1.8 Å or better. The rotamer construction and energy calculations were performed with the XPLOR program.[77] The backbone and C_β atoms of the X-ray structure were fixed during the construction. The sidechain atoms were geometrically constructed from the position of the N, C_α, C, and C_β atoms, using standard bond lengths and bond angles from the CHARMM19 force field. Sidechain dihedral angles were taken from the Tuffery rotamer library.[6]

Additional, solvation energy calculations were performed for the proteins BPTI (4PTI), lysozyme (1LZ1), thioredoxin (2TRX), and ubiquitin (1UBQ). We created a set of 750–1000 random structures for each one by holding the backbone in its native conformation and randomizing the orientation of all sidechains except Pro, Ala, and Cys.

Mean Field Optimization

We did sidechain placement using a mean field approximation.[14] This method calculates iteratively the Boltzmann probability $P(i,k)$ of each rotamer k of each residue i, which is related to the mean energy $E(i,k)$ of sidechain i:

$$E(i,k) = -RT \ln P(i,k). \quad (6)$$

$E(i,k)$ is the Boltzmann average of the interaction energy between sidechain i and its environment; R is the ideal gas constant, and T is the temperature. Since the protein backbone is fixed, we can write

$$E(i,k) = E_{BB}(i,k) + \sum_{j \neq i} \sum_l E(ik, jl) P(j,l) \quad (7)$$

where E_{BB} is the interaction energy with the backbone, the first sum is over protein sidechains j, the second sum is over the rotamers l of sidechain j, and $E(ik,jl)$ is

the interaction energy between sidechains i and j when they occupy rotamers k and l. We assume the optimal sidechain positions correspond to the most probable rotamers.

The probabilities were computed iteratively. The current estimate $P(i,k)^{(n)}$ was updated according to

$$P(i,k)^{(n+1)} = \lambda P(i,k)^{(n)} + (1-\lambda)P(i,k)^{(n-1)}. \quad (8)$$

Different λ values were tested. The best results (reported below) were obtained with $\lambda = 0.35$ and uniform starting rotamer probabilities. About 20 cycles were typically needed for convergence.

An Approximate, Pairwise Surface Area Calculation

In the sidechain placement, the solvent accessible surfaces are calculated by an approximate, but very efficient procedure. The buried surface of a sidechain is computed by summing over the neighboring sidechain and backbone groups. For each neighboring group, the contact area with the sidechain of interest is computed, independently of other surrounding groups. The contact areas are then summed. This approach assumes the contact areas between sidechains are independent, and the total area of a sidechain is a pairwise sum over its neighbors. In fact, surface area buried by one sidechain may also be buried by another. Mayo et al.[81] showed that this approach overestimated the surface areas of buried sidechains, but had little effect on solvent exposed sidechains. Therefore, we performed most of the calculations with a scaling factor applied to the contact areas involving buried sidechains. To determine the optimal scaling factor, following Mayo et al.,[81] we considered the total surface area that is buried within the protein, A_{buried}. A_{buried} is defined as the difference between the total area of all residues, taken separately, and the total surface area of the protein structure; see Ref. 81 for details. With the pairwise approximation, it is computed as

$$A_{\text{buried}} \approx A^{\text{pairwise}}_{\text{buried}} = \sum_i A^0_{i_r t3} - \left(\sum_i A_{i_r t} + \frac{1}{2} \sum_{ij} s_i (A_{i_r j_s t} - A_{i_r t} - A_{j_s t}) \right). \quad (9)$$

The first sum on the right is over all residues i and represents the total area of all residues, taken separately. $A^0_{i_r t3}$ is the exposed area of sidechain i when it occupies the rotamer r, in the presence of just the backbone of residues $i-1$, i, $i+1$ (a tripeptide indicated by the subscript $t3$). The sums in parentheses represent the approximate total protein surface area. $A_{i_r t}$ is the exposed area of the sidechain and its backbone, in the presence of the entire protein backbone. $A_{i_r j_s t}$ is the exposed area of the sidechain pair ij in the presence of the entire protein backbone. Finally, s_i is the scaling parameter, which is set to 1 if residue i is exposed to solvent and to a smaller value $s < 1$ if it is buried. To validate this approximation and determine the optimal s, we performed surface area calculations on the 29 test proteins above, and compared the exact and approximate buried surface areas. Using a scaling factor of $s_i = 0.5$ for buried amino acids [Eq. (9)], we observed a correlation of 0.999 between the approximate and the true surfaces, with a slope of 0.993 and an offset of 73.3 Å^2. The RMSD between the approximate and true surfaces was just 199 Å^2, 1% of the mean surface. Therefore, a value of $s_i = 0.5$ was used for all the sidechain placements, unless otherwise mentioned.

Validation Methods

Two criteria were used to assess the accuracy of the sidechain placements. First, we calculated the percentage of sidechain dihedral angles within 40° of the value in the crystal structure. Second, we computed the RMSD between the sidechain atom positions in the model and the X-ray structure, excluding the C_β. The RMSD calculation took into account the rotational symmetry axis of Asp, Phe, Glu, and Tyr residues.

Later in this article, we compare protocols that use several different effective energy functions, corresponding to different solvent treatments. To determine the significance of the differences between protocols, statistical tests were done. Analysis of variance (ANOVA) was performed to evaluate groups of protocols. The null hypothesis is that for an observable of interest (e.g., RMSD), all the means observed for the various protocols are identical. Rejection of the null hypothesis means that there is at least one pair of means that are different from each other. To identify this pair, we then performed Student's t-tests on all pairs of means.

Charge Mutations

To test the effective energy functions in a situation that mimics protein design calculations, we performed two sets of charge mutations. In the first set, the atomic charges of selected sidechains were modified artificially. These mutations can be classified into three types: (1) charge deletions, which remove the net charge on Arg, Lys, Asp, Glu, and doubly-protonated His; (2) charge insertions, which add a ±1 charge to Ala, Ile, Leu, Val, Met, Pro, Thr, or Tyr; (3) polarity changes, which either make Asn, Gln, or singly-protonated His apolar, or introduce a dipole onto the Cys sidechain. The details of the charge modifications are given in Supplementary Material. While these charge transformations do not correspond to real amino acid mutations, they pose the same physical problems and represent a realistic test of the methodology. They were performed with the CASA, GB, and PE solvent models. Experimental data are not available for this set, so we take PE as a reference, since it is commonly judged to be the most realistic of the three solvent models.[25,29,40] The PE model uses a protein dielectric constant of one, because we wish to model

solvent relaxation, not relaxation of the protein. The CASA model employed the Fraternali atomic solvation parameters:[61] a positive value of 0.0119 kcal/mol/Å^2 for carbon atoms and a negative value of −0.0598 kcal/mol/Å^2 for nitrogen and oxygen atoms. To optimize further the model for charge mutations, we explored several values of the CASA dielectric constant, and we introduced a new solvation parameter for charged atoms, by the following scheme. In charge deletions, the atoms involved have a large charge in their native state and a small one in the mutated state. These atoms were assigned a native-state surface coefficient σ_C (to be determined), and a mutant-state coefficient of 0.0119 kcal/mol/Å^2, reflecting their hydrophobic character after the mutation. The optimum σ_C value was determined by comparison with free energy differences evaluated by Poisson calculations (PE solvent model). In charge insertions, most of the atoms involved are carbons, with a zero initial charge and a large (positive or negative) final charge. For these mutations, the coefficients of all the relevant atoms were set to their chemical-type value at the native state, and to a value of σ_C in the mutant state. In polar-to-neutral mutations, the atoms involved were assigned a native-state coefficient according to their chemical type, and a mutant-state coefficient of −0.0598 kcal/mol/Å^2 (Cys) or 0.0119 kcal/mol/Å^2 (Asn, Gln, His). In all cases, hydrogen atoms were assigned a zero atomic surface coefficient in both states. A list of atom types associated with the new coefficient σ_C is given in Supplementary Material.

The second data set included 140 mutations in 12 proteins for which the experimental stability changes are known (listed in Supplementary Material). In each case, an ionized sidechain is introduced or removed. Each native protein structure is first energy minimized with its backbone fixed. The new sidechain is then positioned successively in each of its rotamers, minimized with the surrounding protein structure fixed, and the best rotamer is retained. To obtain the difference in stability between the mutant and the native protein, the charge modification must also be introduced in the unfolded protein. By subtracting the wildtype/mutant energy differences in the folded and unfolded states, the stability of the mutant relative to the native protein is obtained (see Fig. 1). In our calculations, the "unfolded" state corresponded to the mutated residue, isolated from the rest of the sequence (and with the same coordinates as in the native structure). This simple unfolded model is commonly used in protein design.[65–67] Mutations were performed using the CASA, GB/ACE, GB/HCT, and PE solvent treaments.

Optimization of the GB/HCT Parameters

With the GB/HCT model, each atom is assigned a volume and a scaling factor.[48] For the proteins considered here, there are 29 different atom types in the AMBER protein force field,[72] giving 58 atomic parameters. These were adjusted by an iterative, least-squares procedure to reproduce the stability changes computed with PE for artificial mutations, similar to those described above. 1020 of the mutations were chosen randomly to be optimized; the other 7130 were used for a cross-validation test of the optimization. The PE model employed the AMBER atomic charges and atomic radii optimized very recently by McCammon and coworkers[73] to reproduce a large body of experimental and simulation data. The optimized parameters are given in Supplementary Material.

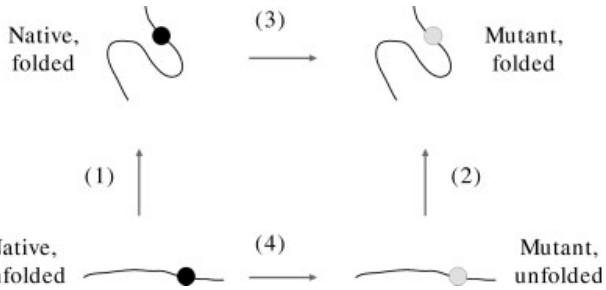


Fig. 1. Thermodynamic cycle used to calculate the change in protein stability due to charge mutations.

RESULTS

Sidechain Placement

We first describe sidechain placement with the CASA solvent model [Eq. (3)]. Twenty-nine proteins were used as a test set (Table III). Several parameterizations were compared in order to optimize the model.

Choice of the dielectric constant

We first considered the effect of the dielectric constant ε and the Coulomb energy term [Eq. (3)], using the CHARMM19 force field and three different sets of atomic solvation parameters, referred to as the Wesson, Ooi, and Fraternali sets, respectively.[59–61] The quality of the predicted structures was assessed by the agreement of sidechain dihedrals with the crystal structure and by the comparison of RMSD of the sidechain atoms with the crystal structure. Results for $\varepsilon = 4$ and $\varepsilon = 20$ are shown in Table I.

The dihedral prediction and sidechain RMSD values are comparable to those of Koehl and Delarue and Yang et al.[14,82] We compared the models with the two ε values and the three solvation parameter sets using the ANOVA statistical test. Differences of χ_1 predictions among the six models are statistically significant at the 1% significance level. The same is true for the predictions of both χ_1 and χ_2 (χ_{1+2}) and for the sidechain RMSD.

Student's test was then done for pairs of models with different ε values. For the Fraternali parameters, differences of the χ_1, χ_{1+2}, and RMSD predictions with the two ε values are not significant at the 1% significance level. The same is true for the Ooi and Wesson parameter sets. Thus, the difference detected by ANOVA is not

TABLE I. CASA Sidechain Predictions with Different Dielectric Constants ε

	Atomic parameters	χ_1 Mean (sd)	χ_{1+2} Mean (sd)	RMSD (Å) Mean (sd)	RMSD, no C_β (Å) Mean (sd)
$\varepsilon = 20$	Fraternali	0.76 (7)	0.61 (6)	1.85 (32)	1.62 (28)
	Wesson	0.71 (7)	0.56 (6)	2.05 (31)	1.77 (26)
	Ooi	0.70 (7)	0.54 (7)	2.12 (37)	1.85 (33)
$\varepsilon = 4$	Fraternali	0.77 (6)	0.63 (6)	1.85 (31)	1.61 (28)
	Wesson	0.72 (6)	0.58 (7)	1.97 (32)	1.73 (29)
	Ooi	0.69 (6)	0.55 (7)	2.13 (32)	1.86 (29)

Fraction of successful dihedral predictions and sidechain RMSD relative to crystal structure, averaged over 29 test proteins. Standard deviation in parentheses (in units of last digit). Rightmost column omits the C_β from RMSD.

TABLE II. CASA Sidechain Predictions with Different Surface Weights α

	Atomic parameters	χ_1 Mean (sd)	χ_{1+2} Mean (sd)	RMSD (Å) Mean (sd)	RMSD, no C_β (Å) Mean (sd)
$\alpha = 1/2$	Fraternali	0.79 (5)	0.66 (6)	1.79 (30)	1.57 (26)
	Wesson	0.76 (6)	0.62 (7)	1.86 (30)	1.63 (27)
	Ooi	0.74 (6)	0.62 (8)	1.94 (34)	1.70 (29)
$\alpha = 1/3$	Fraternali	0.80 (6)	0.67 (6)	1.75 (33)	1.53 (30)
	Wesson	0.78 (6)	0.65 (6)	1.82 (30)	1.59 (27)
	Ooi	0.76 (6)	0.64 (8)	1.88 (30)	1.65 (27)
	Fraternali[a]	0.80 (6)	0.66 (6)	1.74 (33)	1.52 (30)
$\alpha = 0$	none	0.84 (5)	0.68 (6)	1.68 (29)	1.46 (24)

Fraction of successful dihedral predictions and sidechain RMSD relative to crystal structure, averaged over 29 test proteins. Standard deviation in parentheses (in units of last digit). Rightmost column omits the C_β from RMSD. Dielectric constant is $\varepsilon = 4$.
[a] The surface areas of buried sidechains are not downscaled [$s_i = 1$ in Eq. (9)].

related to the choice of ε, which does not affect sidechain prediction. This is consistent with several earlier studies, in which good results were obtained without any electrostatic term.[14,16,21] As pointed out above, however, a model without electrostatics is not appropriate here, since our goal is to do protein design. All the following calculations include the Coulomb energy term with $\varepsilon = 4$.

Choice of the atomic solvation parameters and the weight α

Next, statistical tests were done to determine whether the three solvation parameter sets were significantly different (using $\varepsilon = 4$). Results are summarized in Table II. ANOVA indicates that the null hypothesis is rejected at the 1% significance level for the three quality criteria considered (χ_1, χ_{1+2}, RMSD). Thus, there is at least one prediction protocol that differs. Results with the Ooi and Wesson parameters are equivalent according to the Student test at the 5% significance level for all three quality criteria. However, differences in the χ_1 predictions and the RMSD calculations between Fraternali and the two other parameter sets are statistically significant at the 5% significance level, and differences for χ_{1+2} are significant at the 1% level. Thus, the Fraternali parameters perform better than the other two sets. The Fraternali protocol gave an average success rate for χ_1 and χ_{1+2} predictions of 77% and 63%, respectively.

We next considered a model with $\varepsilon = 4$ but no surface area term [$\alpha = 0$ in Eq. (3)]. The results are actually improved (see Table II). Indeed, the average of the χ_1 and χ_{1+2} predictions are about 84% and 68%, somewhat better than those of Koehl and Delarue and Yang et al.[14,82] It appears that the fully-weighted surface area term is too large, compared to the other terms in the energy function. Nevertheless, further calculations were done with a nonzero α. Indeed, we show below that $\alpha = 1/2$ gives much better results for experimental mutations and their associated stability changes. For protein design, with sidechain reconstruction and mutagenesis occurring simultaneously, we need a consensus model that performs well at both tasks.

Further calculations were therefore done with $\alpha = 1/2$ and $\alpha = 1/3$ (Table II). Results improve with decreasing α. Successful χ_1 prediction with the Fraternali parameters increases to 80% with $\alpha = 1/3$, while χ_{1+2} increases to 67%. These percentages approach those obtained with $\alpha = 0$. The Ooi and Wesson parameters give somewhat poorer results.

We performed ANOVA for each parameter set to compare the different α values (1/3, 1/2, 1). For the Fraternali set, for example, the three protocols were not equivalent according to the χ_1 and χ_{1+2} criteria at the 1% significance level. The $\alpha = 1/2$ and $\alpha = 1/3$ Fraternali models are not significantly different according to the Student's t-test; both are superior to the $\alpha = 1$ model.

Overall, the different Student's tests show that Fraternali with $\alpha = 0$ to 1/2 and Wesson with $\alpha = 1/3$ are equivalent and provide the best results. Table III compares Fraternali, $\alpha = 1/3$ with the work of Koehl and Delarue and Yang et al.[14,82]

TABLE III. CASA Sidechain Predictions Compared to Selected Earlier Work

		CASA, this work (Fraternali, $\varepsilon = 4$, $\alpha = 1/3$)				Koehl and Delarue			Yang et al.		
PDB code	Chain length	χ_1	χ_{1+2}	RMSD[a]	RMSD[b]	χ_1	χ_{1+2}	RMSD[a]	χ_1	χ_{1+2}	RMSD[b]
1PPT	36	0.73	0.64	1.59	1.40	0.73	0.67	1.57	–	–	–
1CRN	46	0.89	0.76	0.88	0.75	0.78	0.68	1.42	0.95	0.84	0.84
2OVO	56	0.90	0.79	1.30	1.12	0.69	0.60	2.35	–	–	–
4PTI	58	0.78	0.70	2.05	1.81	0.87	0.74	1.85	0.80	0.65	1.26
1IGD	61	0.84	0.72	1.22	1.06	–	–	–	0.76	0.74	1.28
1ISU	62	0.95	0.79	1.36	1.19	–	–	–	0.84	0.74	1.12
2SN3	65	0.81	0.68	2.31	2.03	0.59	0.45	2.76	–	–	–
1PTX	68	0.82	0.74	2.16	1.92	–	–	–	0.74	0.61	1.73
1UBQ	76	0.75	0.57	2.19	1.90	0.74	0.57	2.06	–	–	–
1PLC	99	0.77	0.66	1.40	1.21	0.76	0.67	1.55	0.82	0.70	1.24
1LN4	104	0.73	0.62	1.90	1.64	–	–	–	–	–	–
1AAC	105	0.81	0.68	1.81	1.58	–	–	–	0.92	0.69	1.05
256B	106	0.80	0.67	2.01	1.75	–	–	–	0.73	0.53	1.53
2TRX	108	0.82	0.61	1.88	1.65	–	–	–	–	–	–
5CPV	109	0.78	0.61	1.69	1.48	0.68	0.54	1.99	–	–	–
1CCR	112	0.75	0.64	1.86	1.62	–	–	–	0.84	0.60	1.21
1THX	115	0.76	0.65	1.53	1.32	–	–	–	–	–	–
1WHI	122	0.77	0.67	1.92	1.66	–	–	–	0.73	0.65	1.52
1PMY	123	0.80	0.63	1.95	1.70	–	–	–	0.83	0.63	1.23
3RN3	124	0.68	0.60	2.03	1.76	0.66	0.58	2.24	–	–	–
1LZ1	130	0.85	0.70	1.66	1.47	0.80	0.73	1.55	0.82	0.66	1.16
2END	137	0.79	0.63	1.89	1.66	–	–	–	0.79	0.64	1.37
2FOX	138	0.79	0.69	1.46	1.27	–	–	–	0.74	0.56	1.35
2HBG	147	0.77	0.61	1.79	1.57	–	–	–	0.76	0.61	1.34
2RN2	155	0.81	0.67	2.01	1.77	–	–	–	–	–	–
2CPL	165	0.86	0.71	1.49	1.31	–	–	–	–	–	–
1KOE	172	0.73	0.63	1.85	1.61	–	–	–	–	–	–
1XNB	185	0.81	0.70	1.58	1.39	–	–	–	0.78	0.66	1.89
1ES9	212	0.78	0.61	1.92	1.68	–	–	–	–	–	–
Mean	110	0.80	0.67	1.74	1.53	0.72	0.62	1.89	0.80	0.66	1.36

Fraction of successful dihedral predictions and sidechain RMSD (Å) relative to crystal structure.
[a] C_β excluded.
[b] C_β included.

To test for a possible force-field dependency of these results, calculations with the AMBER force field[72] were performed for six of the proteins. The accuracy of sidechain prediction is equivalent to that observed with the CHARMM19 force field, with the $\varepsilon = 4$, $\alpha = 1/3$, Fraternali protocol (data not shown). For example, the average χ_1 prediction rate is 75% with both AMBER and CHARMM19 for these six proteins.

The rate of successful prediction was analyzed as a function of amino acid type. In agreement with earlier work, it is easier to predict the rotamers of hydrophobic residues. Their rate of successful χ_1 prediction is 92% with the best parameterization, compared to 71% for hydrophilic residues. This can be explained because hydrophobic residues are mainly located in the protein core and are constrained by van der Waals packing interactions. There is little difference in accuracy between large and small residues. A low χ_{1+2} accuracy is observed for Asn, Gln, and His, since these residues' χ_2 angle can often be flipped by 180° with only a small change in energy.

Solvation Energies: Comparing Solvent Treatments

To directly compare the three solvent models, CASA, GB, and PE, we computed Coulomb and solvation energies for four proteins: trypsin inhibitor (4PTI), thioredoxin (2TRX), lysozyme (1LZ1), and ubiquitin (1UBQ). For each protein, we generated 750–1000 structures by randomizing sidechain rotamers. In the CASA energies, the Coulomb term is divided by a dielectric constant ε [Eq. (3)]. We initially searched a range of values, $\varepsilon = 1$–20, to find the optimum for each protein.

The RMSDs between PE and CASA are listed in Table IV. The optimal dielectric constants are fairly uniform among proteins: $\varepsilon = 1.5$–2.5. With these ε values, the PE/CASA RMSD ranges from 33 kcal/mol (4PTI) to 58 kcal/mol (2TRX). Ultimately, we require a consensus CASA model that performs well for sidechain placement, solvation energies, and charge mutations, and for all proteins. Sidechain placement (above) is insensitive to ε, while charge mutations (below) work best with a large value, $\varepsilon = 16$–20. Table IV also reports the CASA solva-

TABLE IV. Protein Solvation: RMS Deviation Between the CASA and GB Energies and the PE Reference Values

	CASA dielectric constant ε								
Protein	1	1.5	2.0	2.5	3.0	4.0	20.0	GB/ACE	GB/HCT
4PTI	46.9	**32.5**	32.8	35.7	38.5	42.9	55.6	71.7	20.7
2TRX	87.7	64.8	59.1	**58.3**	59.0	61.1	71.0	99.6	30.8
1LZ1	80.6	59.1	**56.2**	57.7	60.0	64.1	78.3	163.5	44.7
1UBQ	70.1	52.0	**49.8**	51.3	53.4	57.1	69.6	102.1	26.4

All values in kcal/mol. For each protein, the deviations are evaluated over 750–1000 rotameric structures. Best CASA results in boldface.

TABLE V. Comparison Between the CASA and Poisson Models for Charge Mutations

Protein	No of mutations	Optimal σ_C (kcal/mol/Å^2)	Optimal ε	PE-CASA RMSD[a] (kcal/mol)	Success[b] rate (%)
AspRS	518	−0.15 (−0.20)	16	17.5 (17.9)	80.1 (80.1)
3RN3	106	−0.20	16	13.1	79.0
4PTI	51	−0.20	16	11.0	74.4
5CYT	84	−0.20	16	13.5	79.8
2TRX	94	−0.20	16	12.3	80.8
1LZ1	108	−0.20	16	12.2	83.4
1UBQ	67	−0.20	16	10.7	86.5

Proteins are indicated by their PDB code, except AspRS, which is *Escherichia coli* aspartyl-tRNA synthetase. The surface coefficient σ_C is associated with atoms on charged sidechains (see text).
[a]RMS deviation between the surface area (CASA) and Poisson (PE) energy differences.
[b]Percentage of mutations that are predicted to have a positive or negative stability change by both CASA and PE.

tion results with $\varepsilon = 20$. The RMS deviations from PE increase to 56—71 kcal/mol. Thus, the loss in performance is significant when one takes the charge mutation dielectric as a consensus value for all proteins.

The CASA, GB/ACE, and GB/HCT total solution energies are compared with PE in Figure 2. Note that CASA and GB/ACE use the Charmm19 atomic charges, and so they are compared with PE performed with these charges. GB/HCT uses the AMBER charges, and is compared with PE performed with the AMBER charges. A CASA ε of 20 is used. There is a considerable energy dependency on the rotameric structure, with values between about −300 and +300 kcal/mol. The trend of the PE energies is reproduced approximately by both GB/ACE and CASA. The CASA surface term is almost constant for all structures (not shown), so that the CASA behavior is governed by the Coulomb term [Eq. (3)]. The CASA/PE correlation indicates, therefore, that the PE solvation energies correlate quite well with the Coulomb energy. The best correlation by far is obtained with GB/HCT. The RMS deviation between GB/HCT and PE is also quite small (21–45 kcal/mol; Table IV), considerably lower than that of CASA with the consensus dielectric constant.

Charge Modifications: Comparing Solvent Treatments

The third and most important test of the solvent models is the calculation of stability changes due to mutations that introduce or remove charged groups. In protein design, mutations are introduced randomly, and the largest stability changes will usually be associated with changes in the net protein charge. This is especially true for mutations that affect charges in fully or partly buried positions.[69] We performed two sets of mutations. The first included over 1000 mutations in seven proteins for which experimental data are not available, described in this section. The second included 140 mutations in 12 proteins for which experimental data are available, described in the next section.

Artificial mutations: CASA vs. PE

We used the Fraternali atomic surface parameters, along with a new atomic parameter σ_C for certain atoms belonging to ionized groups. The new parameter allows us to optimize the accuracy of the CASA model for charge deletions and insertions. The atom types associated with this new coefficient are listed in Supplementary Material. Values of the atomic parameter σ_C between −0.20 kcal/mol/Å^2 and the original value, −0.0598 kcal/mol/Å^2, were tried, in combination with dielectric constants in the range $\varepsilon = 2$–20. The CASA and PE energy changes for each mutation were compared. The combination of σ_C and ε that maximizes the agreement between CASA and PE is listed in Table V for the seven proteins. The optimum coefficient for atoms on charged sidechains (in their charged state) is between −0.15 and −0.20 kcal/mol/Å^2, and the optimum dielectric constant is 16–20. With these values, the RMSD between the CASA and PE stability changes ranges from 17.5 kcal/mol for the largest protein (AspRS) to 10.0–11.0 kcal/mol for the smallest proteins (ubiquitin and BPTI).

As explained above, we would like to obtain a consensus parameterization that works well for all proteins. The best σ_C and ε values are the same for all but one protein: a slightly smaller σ_C of −0.15 kcal/mol/Å^2 is

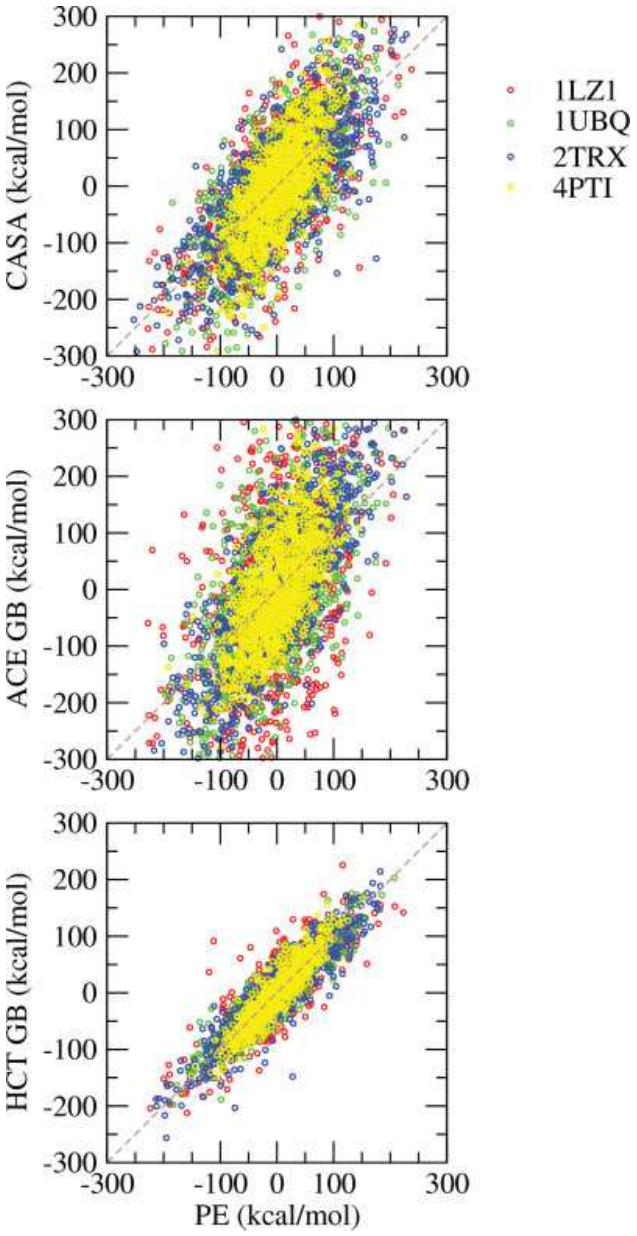


Fig. 2. Total solution energies of protein rotamer structures with the GB/ACE, surface area (CASA), and GB/HCT models, versus the corresponding Poisson values (PE). Points are colored according to the protein, listed by their PDB codes: 1LZ1, lysozyme; 1UBQ, ubiquitin; 2TRX, thioredoxin; 4PTI, bovine pancreatic trypsin inhibitor.

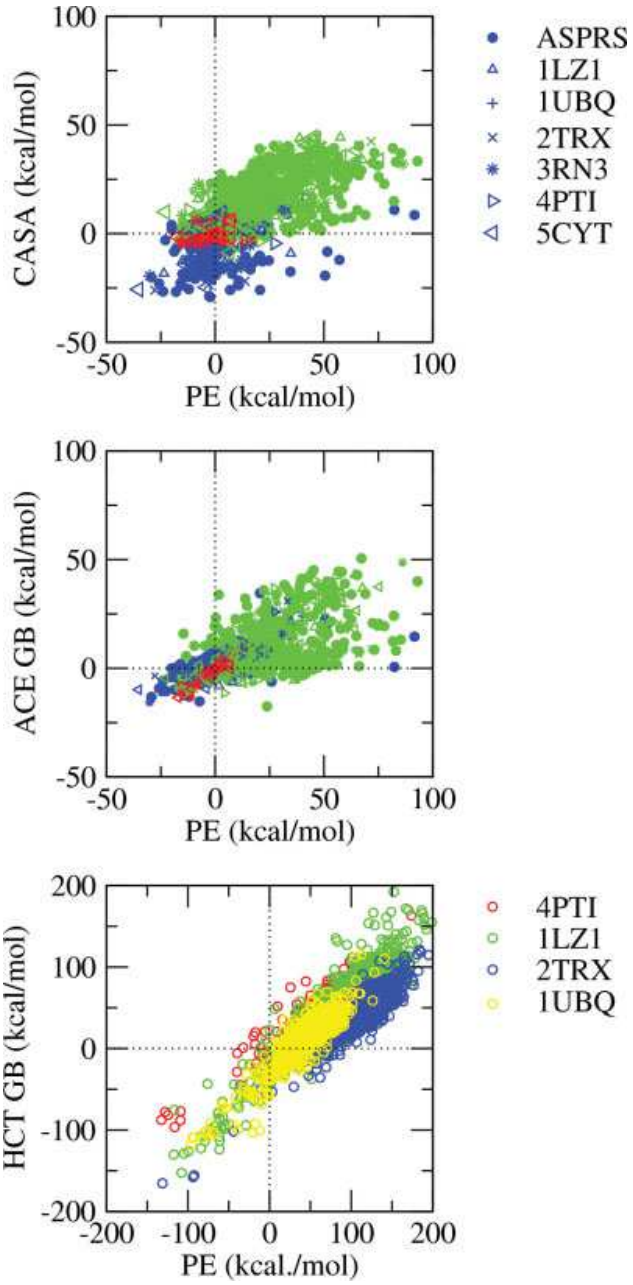


Fig. 3. Stability changes due to charge mutations with CASA, GB/ACE, and GB/HCT, vs. the corresponding changes in the PE model. For CASA and GB/ACE, green, blue and red points correspond, respectively, to charge insertions, charge deletions and polar-to-neutral conversions; different symbols correspond to different proteins, listed by their PDB codes. For GB/HCT, points are colored by protein.

best for AspRS. Using the consensus σ_C value for AspRS gives almost the same result (Table V). The same σ_C and ε values work well for the experimental mutations, explained later. We conclude that they are likely to work well for most mutations in most proteins.

The CASA stability changes are plotted against the PE ones in Figure 3. Most charge insertions decrease the protein stability, both in the CASA and PE models (i.e., they give a positive double free energy difference, $\Delta G_{mut} - \Delta G_{nat}$, for the thermodynamic cycle in Fig. 1). On the other hand, most charge deletions increase the protein stability. Most charged residues are at the protein surface and their interactions with the rest of the protein are screened by the nearby solvent. Eliminating the charge on such residues destabilizes the folded state, mainly due to the loss of interactions between the charge and the solvent. In the unfolded state, the same charges

TABLE VI. Comparison Between the GB and Poisson Models for Charge Mutations

		GB/ACE, Approx 1		GB/ACE, Approx 2		GB/HCT	
Protein	No of mutations	RMSD[a]	% success[b]	RMSD[a]	% success[b]	RMSD[a]	% success[b]
AspRS	518	40.7	47.9	18.2	82.2		
3RN3	106	21.6	66.9	14.0	84.0		
4PTI	51	13.7	74.5	7.0	70.6	8.3	99.2
5CYT	84	15.2	65.5	12.6	73.8		
2TRX	94	26.3	64.9	17.4	91.5	12.1	98.5
1LZ1	108	22.8	59.2	12.3	80.6	16.5	98.2
1UBQ	67	17.1	56.8	11.6	76.1	10.8	92.3

[a]RMS deviation between the GB and Poisson (PE) total energy differences (kcal/mol).
[b]Percentage of mutations predicted to have a positive or negative stability change by both GB and PE. Approximations 1 and 2 are explained in the text. Proteins are indicated by their PDB code, except AspRS, which is *Escherichia coli* aspartyl-tRNA synthetase.

are even more exposed to solvent. Thus, the charge elimination destabilizes the unfolded state even more, yielding a negative (favorable) double free energy difference. CASA and PE produce the same stability order (i.e., sign of the stability change) for 81% of mutations in all proteins. The CASA energy differences vary in a range of ≈80 kcal/mol, whereas the corresponding PE range is 150 kcal/mol; nevertheless, the two distributions are reasonably well-correlated.

Calculations were also done with a "distance-dependent" dielectric constant, $\varepsilon(r) = \varepsilon_0 r$, where r is the separation between a pair of charges and ε_0 is a constant. The success rate increased to about 86% if one uses $\varepsilon_0 = 2$. However, the performance of this model variant is much poorer for the experimental mutations, below. Increasing ε_0 improves the experimental mutations but deteriorates the artificial ones (not shown).

Artificial mutations: GB/ACE and GB/HCT vs. PE

The mutations are of three types: (1) charge insertions (usually, introduction of charge onto hydrophobic atom types); (2) charge deletions (elimination of charge on hydrophilic atom types), and (3) polar-to-neutral conversions. Thus, the mutated atoms have a combination of charge and van der Waals radius that is different from the optimum value for GB calculations. For this reason, it was necessary to modify some of the GB/ACE parameters. Table VI contains the results of two approximations. The first approximation corresponds to the usual GB/ACE parameterization (i.e., V01 volumes scaled by 0.8).[52] In the second approximation, the hydrogen types HC are assigned a Voronoi volume of 4.0, the V01 volumes of atom types CH1E, CH2E, CH3E (carbons of hydrophobic sidechains, carrying a significant partial charge in the charge-insertion mutations) were scaled by a factor 0.4, and nitrogen types NH3, NC2 were set to the original V01 volumes; for all other atom types, the V01 volumes were scaled by 0.8. The same parameterization is used for all proteins. As seen from Table VI, approximation 2 yields an RMSD between GB/ACE and PE of 7–18 kcal/mol. The fraction of mutations predicted with a positive or negative stability change by models is 80%. Thus, with this set of parameters, the GB/ACE model is comparable to CASA. Another recent study reported poorer results with GB/ACE for a similar test.[20] This was presumably due to the use of an early and nonoptimal parameterization of GB/ACE; see Calimet et al.[52] for a critical discussion of the early parameterization. The GB/ACE stability changes with approximation 2 are plotted against the corresponding PE results in Figure 3. Most charge insertions destabilize the protein, as with CASA. On the other hand, most charge deletions increase the stability.

The GB/HCT results are also given in Table VI and Figure 3. The GB/HCT parameters were extensively fitted to the PE model in this work (see Methods section). However, all the GB/HCT—PE comparisons correspond to cross validation tests, using PE data not employed during the fits. The set of proteins and mutations is slightly different from GB/ACE, with fewer (four) proteins, but a larger number of mutations per protein. The RMSD between GB/HCT and PE is 8—16 kcal/mol, smaller than for CASA and GB/ACE. Correlation between GB/HCT and PE is excellent, far superior to CASA and GB/ACE, with 97% of the stability changes predicted to have the correct sign.

Charge Modifications: Comparing to Experiment

We considered 140 mutations in 12 proteins (Table VII), with experimental stability changes taken either from the ProTherm database[83] or Ref. 74. All mutations introduced or removed an ionized residue. The stability changes computed with CASA, GB/ACE, GB/HCT, and PE are summarized in Table VII and detailed in Supplementary Material. We report the mean unsigned error (MUE) for each protein, with selected outliers left out. The outliers were identified by a large van der Waals contribution to the stability change (10 kcal/mol or more). Such a large contribution reflects steric conflict in the mutant structure that results from our simple sidechain construction method. Overall, we identify 7 outliers with GB/HCT, 8 with CASA, and 21 with GB/ACE. Calculations with GB are slower than with CASA; e.g., with GB/HCT, a typical

TABLE VII. Comparison Between Models and Experiment for Charge Mutations

Protein	PDB code	Total number of mutants	Mean unsigned error (kcal/mol)			
			CASA $\varepsilon = 16$	GB/ACE $\varepsilon_P = 4$	GB/HCT $\varepsilon_P = 4$	PE $\varepsilon_P = 4$
Protein G, B_1 domain	1EM7	2	0.31 (2)	3.5 (2)	2.7 (2)	1.6 (2)
Chymotrypsin inhibitor	1YPC	15	3.3 (15)	4.3 (13)	4.4 (15)	4.6 (15)
Lysozyme	2LZM	14	2.6 (13)	2.6 (12)	4.2 (14)	4.4 (14)
Ribonuclease	2RN2	25	2.5 (24)	3.4 (23)	3.6 (24)	4.8 (24)
Src SH3 domain	1SHG	5	1.8 (5)	4.6 (5)	3.6 (5)	3.2 (5)
Staphylococcal nuclease	1STN	45	1.9 (40)	2.2 (33)	3.8 (42)	3.9 (42)
Thioredoxin	2TRX	3	5.1 (3)	6.8 (2)	1.4 (2)	1.8 (2)
Trypsine	1BPI	11	1.4 (10)	2.4 (10)	1.7 (9)	6.7 (9)
Ubiquitin	1UBQ	8	2.1 (8)	1.5 (8)	1.6 (8)	1.4 (8)
pepT1[c]	–	4	1.6 (4)	2.4 (4)	2.8 (4)	1.9 (4)
K2AE2[c]	–	4	1.9 (4)	2.3 (4)	1.8 (4)	1.4 (4)
KEAKE[c]	–	4	1.8 (4)	2.0 (4)	1.6 (4)	1.2 (4)
Total		140	2.1 (132)[a]	2.9 (120)	3.4 (133)	3.9 (133)
					2.1[b] ($\varepsilon_P = 8$)	2.6[b] ($\varepsilon_P = 8$)

[a]In parentheses: the number of mutations after discarding those with a van der Waals contribution of 10 kcal/mol or more.
[b]GB/HCT or PE supplemented by a surface area term and using a protein dielectric of 8 (see text).
[c]Experimental data from Ref. 88. Natives structures are not known experimentally; they were therefore modelled using the Swiss PDB Viewer, which constructs an ideal α-helix and places sidechains in favorable rotamers.[89]

mutation takes 2–4 times more CPU time than with CASA.

Excluding the outliers, the MUE is 2.1 kcal/mol with CASA. Poorer results are obtained if the constant dielectric in the Coulomb electrostatic term is replaced by a distance-dependent dielectric, $\varepsilon(r) = \varepsilon_0 r$: with $\varepsilon_0 = 2$, for example, the error increases to 7.9 kcal/mol. Similarly, while sidechain reconstruction (above) works well without any surface term [$\alpha = 0$ in Eq. (3)], results here are much poorer: the MUE increases to 4.6 kcal/mol if the surface term is left out. These results illustrate further that the best model for protein design should be a consensus model, not necessarily optimal for each prediction task but reasonably competent for all of them.

In the case of GB/ACE, GB/HCT, and PE, the protein dielectric constant ε_P is adjusted empirically to minimize the MUE. Indeed, our simple sidechain modeling does not allow dielectric relaxation of the protein structure when an ionized sidechain is introduced or removed. For all three models, GB/ACE, GB/HCT, and PE, the best results are obtained with a protein dielectric of 3–5. Note that the use of GB models with a protein dielectric greater than 1 [Eq. (4)] is not very common, but is straightforward.[42,84] A consensus value of four for the protein dielectric works well with all three models, giving MUEs of 3.4 kcal/mol with GB/HCT and 3.9 kcal/mol with PE. Results with a dielectric of one are very poor (e.g., the MUE is 8.7 kcal/mol with PE). With GB/ACE, mutations involving Asp and Glu sidechains were found to be poorly described. Therefore, another adjustable parameter was introduced, which empirically increases by 6.3 kcal/mol the contribution of Asp or Glu sidechains to the stability of the unfolded state. This leads to a MUE of 2.9 kcal/mol with GB/ACE (using $\varepsilon_P = 4$, and with 21 outliers omitted). With GB/HCT and PE, the prediction quality did not depend noticeably on the amino acid type.

Mutation of a charged sidechain could, in some cases, alter the protonation state of surrounding residues. It is not practical to model this effect in detail, because there are too many possible protonation states to consider. It is implicitly incorporated into the model through the choice of a protein dielectric constant greater than 1. Furthermore, for each protein, we systematically did CASA calculations with the histidine sidechains either all ionized or all neutral. Results for the mutations were reasonably similar (not shown).

The performance of the three models, GB/ACE, GB/HCT, and PE, could be improved by adding an additional, surface area term to describe hydrophobic solvation and dispersion interactions with the surrounding solvent, as is commonly done in protein modeling, see e.g. Refs. 32,85. Using a surface coefficient of $\sigma = -0.04$ kcal/mol/Å^2 for all atoms and a somewhat larger dielectric of $\varepsilon = 8$, the mean unsigned error for GB/HCT drops to 2.1 kcal/mol, identical to the CASA value. For PE, the mue drops to 2.6 kcal/mol, using a surface coefficient of $\sigma = -0.05$ kcal/mol/Å^2 for all atoms and a dielectric of $\varepsilon = 8$. We did not try this procedure with GB/ACE, because of its poorer overall performance.

The experimental agreement is slightly worse with PE than with CASA and GB/HCT. Nevertheless, we took PE to be the reference model for the artifical mutations, above. Indeed, it is the model with the clearest physical basis and the best performance in general for protein electrostatics.[29,68] CASA, in contrast, is not expected to give quantitative accuracy for mutations of buried sidechains, since the surface energy cannot distinguish between positions that are deeply buried and positions that are closer to the protein surface. GB/HCT has a MUE close to PE, largely because its parameters have been optimized to reproduce PE. Thus, even though the PE mue is slightly larger, our results do not contradict

many earlier studies of protein electrostatics showing that PE is a valid reference model.

CONCLUSIONS

We have examined three problems that are important ingredients of computational protein design: sidechain placement, protein solvation, and mutagenesis involving charged sidechains. We have tested the behavior of four implicit solvent models: the Poisson model, considered to be the standard of accuracy, the GB/ACE and GB/HCT generalized Born models, and the CASA model. The CASA model is commonly used for protein design; GB/ACE and GB/HCT are more recent and sophisticated solvent models. Several methods have been proposed that allow GB models to be used for protein design.[19,23,38] The most recent one achieves a residue–pairwise GB implementation without any loss in accuracy.[23]

Using a standard mean field method for sidechain placement along with the CASA solvent model, we obtain results of similar quality to earlier workers.[14,82] The results are weakly sensitive to the details of the CASA model: solvent dielectric, overall weight of the surface term, force field. This is consistent with the good results obtained by some earlier workers without any electrostatic term or solvent model.[14,16,21]

In contrast, the mutagenesis results are much more sensitive to the solvent treatment. This is expected, since almost all the mutations involved insertions or deletions of a net charge. Many of the insertions were on buried sidechains. All four solvent models were compared. Different kinds of parameter optimization were performed. In the CASA model, we introduced a new atomic surface parameter for oxygen and nitrogen atoms in charged sidechains and we adjusted the dielectric constant. In the GB/ACE model, we adjusted the atomic volumes of selected atom types belonging to charged sidechains. In the GB/HCT model, we optimized the atomic volumes and scaling factors (58 parameters in all). All these optimizations used a large data set of artifical mutations, and took the Poisson model as the reference. The Poisson model employed a dielectric constant of one for the protein (and 80 for solvent), because we are optimizing the implicit solvent treatments, which try to reproduce the relaxation of solvent, not protein, in response to the mutations. Similarly, the GB models employed a protein dielectric of one and a solvent dielectric of 80. CASA uses a single dielectric constant, which tries to capture the average effect that solvent relaxation exerts within the protein interior. A value of about 20 worked well—intermediate between the solvent value of 80 and the value of one that is appropriate for the protein interior in these calculations. With these parameter adjustments and dielectric constants, the quality of CASA and GB/ACE were very similar, as compared to the Poisson reference. In 80–81% of the charged mutations, they correctly captured the sign and order of magnitude of the protein stability change. With GB/HCT, the sign was correct for 97% of the mutations, a dramatic improvement with this more recent GB variant.[48] Protein solvation energies were also strikingly better with GB/HCT than with CASA or GB/ACE.

Finally, in a separate test, we compared the models to experiment for a set of 140 point mutations. Comparison with experiment introduces several major difficulties. First, the stability changes are very small, so that a simple null model will usually give better agreement than any current implicit solvent model. Second, we need a model for the unfolded protein, whose structure is not known. Third, we must describe the structural relaxation *of the protein*, and not just that of the solvent. Fourth, we must describe hydrophobic contributions to stability, which are notoriously hard to capture with simple models.[25,86] To describe the unfolded state, we adopted the simplest possible, commonly-used, tripeptide model. To describe the protein relaxation, we increased the protein dielectric in the GB and PE models, exploring values between 2 and 32. With CASA, we kept a dielectric of 20, and verified that increasing it had only a small effect on the results. With GB/ACE, we also added an empirical correction for Asp/Glu sidechains in the unfolded state. This correction presumably reflects a problem with the GB/ACE parameterization of carboxylate groups. Finally, we incorporated hydrophobic contributions, along with solvent dispersion interactions into the GB/HCT and PE models by adding a surface term. All the solvent models gave fair agreement with experiment, with mean unsigned errors of 2.1 kcal/mol for CASA, 2.9 kcal/mol for GB/ACE, and 2.1 kcal/mol for GB/HCT supplemented by the surface term. GB/HCT also gave excellent results for the artifical mutations and the solvation energies.

In summary, (a) we confirm earlier observations that the choice of electrostatic model is not very important for sidechain placement; (b) GB/HCT and CASA give the same accuracy for surface mutations; (c) GB/HCT yields an enormous improvement for total solvation free energies and for mutations in fully or partly buried positions. Thus, for problems like protein design that involve all these aspects, the most recent GB models, and their best pairwise implementations,[87] represent an important step forward.

ACKNOWLEDGMENTS

We thank A. Jaramillo and M. Schmidt am Busch for useful discussions. Support was provided by the Egide program ZENON between Cyprus and France (to GA and TS), by the ACI IMPBio Program of the French Ministry of Research (to TS), by the HPC-EUROPA project RII3-CT-2003-506079 (to GA and TS), and by a Cyprus "Research Reinforcement of Young Cypriot Researchers" (PENEK) grant (to GA).

REFERENCES

1. Hellinga H. Metalloprotein design. Curr Opin Biotech 1996;7: 437–441.
2. Al-Lazikani B, Jung J, Xiang Z, Honig B. Protein structure prediction. Curr Opin Chem Biol 2001;5:51–56.
3. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.

4. Bolon D, Mayo S. Enzyme-like proteins by computational design. Proc Natl Acad Sci USA 2001;98:14274–14279.
5. Ponder J, Richards FM. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1988;193:775–791.
6. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267.
7. Dunbrack R, Karplus M. Backbone-dependent rotamer library for proteins. Application to sidechain prediction. J Mol Biol 1993; 230:543–574.
8. Dunbrack R, Cohen F. Bayesian statistical analysis of protein sidechain rotamer preferences. Prot Sci 1997;6:1661–1681.
9. Desmet J, De Mayaer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein sidechain positioning. Nature 1992;356:539–542.
10. Goldstein R. Efficient rotamer elimination applied to protein sidechains and related spin glasses. Biophys J 1994;66:1335–1340.
11. Looger L, Hellinga H. Generalized dead-end elimination algorithms make large-scale protein sidechain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol 2001;307:429–445.
12. Holm L, Sander C. Database algorithm for generating protein backbone and sidechain coordinate from a C_α trace: application to model building and detection of coordinates errors. J Mol Biol 1991;218:183–194.
13. Lee C, Subbiah S. Prediction of protein sidechain conformation by packing optimization. J Mol Biol 1991;217:373–388.
14. Koehl P, Delarue M. Application of a self-consistent mean field theory to predict protein sidechain conformations and estimate their conformational entropy. J Mol Biol 1994;239:249–275.
15. Mendes J, Baptista A, Carrondo M, Soares C. Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants. J Computer Aided Mol Des 2001;15:721–740.
16. Liang S, Grishin N. Sidechain modeling with an optimized scoring function. Prot Sci 2002;11:322–331.
17. Onufriev A, Bashford D, Case D. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. Proteins 2004;55:383–394.
18. Feig M, Onufriev A, Lee M, Im W, Case D, Brooks CL, III. Performance comparison of generalized Born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem 2004;25:265–284.
19. Pokola N, Handel T. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. Prot Sci 2004;13:925–936.
20. Jaramillo A, Wodak S. Computational protein design is a challenge for implicit solvation models. Biophys J 2005;88:156–171.
21. Canutescu AA, Shelenkov AA, Dunbrack R. A graph-theory algorithm for rapid protein side-chain prediction. Prot Sci 2003; 12:2001–2014.
22. Feig M, Brooks CL, III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr Opin Struct Biol 2004;14:217–224.
23. Archontis G, Simonson T. A residue-pairwise Generalized Born scheme suitable for protein design calculations. J Phys Chem B 2005;109:22667–22673.
24. Tanford C. The hydrophobic effect. New York: John Wiley; 1980.
25. Roux B, Simonson T. Implicit solvent models. Biophys Chem 1999;78:1–20.
26. Warwicker J, Watson H. Calculation of the electrostatic potential in the active site cleft due to α helix dipoles. J Mol Biol 1982;157:671–679.
27. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
28. Schaefer M, Vlijmen Hv, Karplus M. Electrostatic contributions to molecular free energies in solution. Adv Prot Chem 1998;51:1–57.
29. Simonson T. Electrostatics and dynamics of proteins. Rep Prog Phys 2003;66:737–787.
30. Simonson T. Macromolecular electrostatics: continuum models and their growing pains. Curr Opin Struct Biol 2001;11:243–252.
31. Sitkoff D, Sharp K, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.
32. Simonson T, Brünger AT. Solvation free energies estimated from macroscopic continuum theory: an accuracy assessment. J Phys Chem 1994;98:4683–4694.
33. Bashford D, Karplus M. The pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry 1990;29:10219–10225.
34. Archontis G, Simonson T, Karplus M. Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase. J Mol Biol 2001;306:307–327.
35. Hendsch Z, Tidor B. Electrostatic interactions in the GCN4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding. Prot Sci 1999;8:1381–1392.
36. Gohlke H, Kiel C, Case D. Insight into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. J Mol Biol 2003;330:891–913.
37. David L, Luo R, Gilson M. Comparison of Generalized Born and Poisson models: energetics and dynamics of HIV protease. J Comput Chem 2000;21:295–309.
38. Wisz M, Hellinga H. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins 2003;51:360–377.
39. Still WC, Tempczyk A, Hawley R, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
40. Bashford D, Case D. Generalized Born models of macromolecular solvation effects. Ann Rev Phys Chem 2000;51:129–152.
41. Hawkins G, Cramer C, Truhlar D. Pairwise descreening of solute charges from a dielectric medium. Chem Phys Lett 1995;246:122–129.
42. Schaefer M, Karplus M. A comprehensive analytical treatment of continuum electrostatics. J Phys Chem 1996;100:1578–1599.
43. Qiu D, Shenkin P, Hollinger F, Still W. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J Phys Chem A 1997;101:3005–3014.
44. Ghosh A, Rapp C, Friesner RA. Generalized Born model based on a surface area formulation. J Phys Chem B 1998;102:10983–10990.
45. Dominy B, Brooks CL, III. Development of a Generalized Born model parameterization for proteins and nucleic acids. J Phys Chem B 1999;103:3765–3773.
46. Lee M, Salsbury Jr, F, Brooks CL, III. Novel generalized Born methods. J Chem Phys 2002;116:10606–10614.
47. Wagner F, Simonson T. Implicit solvent models: combining an analytical formulation of continuum electrostatics with simple models of the hydrophobic effect. J Comput Chem 1999;20:322–335.
48. Onufriev A, Bashford D, Case D. Modification of the generalized Born model suitable for macromolecules. J Phys Chem B 2000; 104:3712–3720.
49. Simonson T, Carlsson J, Case DA. Proton binding to proteins: pK_a calculations with explicit and implicit solvent models. J Am Chem Soc 2004;126:4167–4180.
50. Lee M, Salsbury F, Jr., Brooks C, III. Constant pH molecular dynamics using continuous titration coordinates. Proteins 2004;56: 738–752.
51. Cornell W, Abseher R, Nilges M, Case D. Continuum solvent molecular dynamics study of flexibility in interleukin-8. J Mol Graph Model 2001;19:136–145.
52. Calimet N, Schaefer M, Simonson T. Protein molecular dynamics with the generalized Born/ACE solvent model. Proteins 2001; 45:144–158.
53. Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A. Exhaustive docking of molecular fragments with electrostatic solvation. Proteins 1999;37:88–105.
54. Bursulaya B, Brooks C, III. Comparative study of the folding free energy landscape of a three-stranded β-sheet protein with explicit and implicit solvent models. J Phys Chem B 2001;104: 12378–12383.
55. Simmerling C, Strockbine B, Roitberg A. All-atom structure prediction and folding simulations of a stable protein. J Am Chem Soc 2002;124:11258–11259.
56. Rapp C, Friesner R. Prediction of loop geometry using a generalized Born model of solvation effects. Proteins 1999;35:173–183.
57. Felts A, Gallicchio E, Wallqvist A, Levy R. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized Born solvent model. Proteins 2002;48:404–422.

58. Eisenberg D, McClachlan A. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
59. Ooi T, Oobatake M, Nemethy G, Scheraga H. Accessible surface areas as a measure of the thermodynamic hydration parameters of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.
60. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Prot Sci 1992;1:227–235.
61. Fraternali F, van Gunsteren W. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. J Mol Biol 1996;256:939–948.
62. Ferrara P, Apostolakis J, Caflisch A. Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins 2002;46:24–33.
63. Pei J, Wang Q, Zhou J, Lai L. Estimating protein-ligand binding free energy: atomic solvation parameters for partition coefficient and solvation free energy calculation. Proteins 2004; 57:661–664.
64. Dahiyat B, Mayo S. Protein design automation. Prot Sci 1996;5: 895–903.
65. Koehl P, Levitt M. Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 2002;99: 1280–1285.
66. Ogata K, Jaramillo A, Cohen W, Briand J, Conan F, Wodak S. Automatic sequence design of MHC class-I binding peptides impairing CD8+ T cell recognition. J Biol Chem 2003;278:1281.
67. Liang S, Grishin N. Effective scoring function for protein sequence design. Proteins 2004;54:271–281.
68. Baker N. Poisson-Boltzmann methods for biomolecular electrostatics. Methods Enzym. 2004;383:94.
69. Dwyer J, Gittis A, Karp D, Lattman E, Spencer D, Stites W, Garcia-Moreno B. High apparent dielectric constants in the interior of a protein reflect water penetration. Biophys J 2000;79: 1610–1620.
70. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. Charmm: a program for macromolecular energy, minimization, and molecular dynamics calculations. J Comput Chem 1983;4:187–217.
71. Hawkins G, Cramer C, Truhlar D. J Phys Chem 1996;100:19824.
72. Cornell W, Cieplak P, Bayly C, Gould I, Merz K, Ferguson D, Spellmeyer D, Fox T, Caldwell J, Kollman P. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
73. Swanson J, Adcock S, McCammon J. Optimized radii for Poisson-Boltzmann calculations with the AMBER force field. J Chem Theory Comput 2005;1:484–493.
74. Guérois R, Nielsen J, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.
75. Simonson T, Archontis G, Karplus M. Free energy simulations come of age: the protein–ligand recognition problem. Acc Chem Res 2002;35:430–437.
76. Lee B, Richards F. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
77. Brünger AT. X-plor version 3.1, A system for X-ray crystallography and NMR. New Haven: Yale University Press; 1992.
78. Schaefer M, Bartels C, Leclerc F, Karplus M. Effective atom volumes for implicit solvent models: comparison between Voronoi volumes and minimum fluctuation volumes. J Comp Chem 2001; 22:1857–1879.
79. Moulinier L, Case D, Simonson T. X-ray structure refinement of proteins with the generalized Born solvent model. Acta Cryst D 2003;59:2094–2103.
80. Madura J, Briggs J, Wade R, Davis M, Luty B, Ilin A, Antosiewicz J, Gilson M, Baheri B, Scott L, McCammon J. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. Comput Phys Commun 1995;91:57–95.
81. Street A, Mayo S. Pairwise calculation of protein solvent-accessible surface areas. Folding Des 1998;3:253–258.
82. Yang J, Tsai C, Hwang M, Tsai H, Hwang J, Kao C. GEM: a Gaussian evolutionary method for predicting protein sidechain conformations. Prot Sci 2002;11:1897–1907.
83. Kumar M, Bava K, Gromiha M, Parabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucl Acids Res 2006;34:D204–206.
84. Sigalov G, Scheffel P, Onufriev A. Incorporating variable dielectric environments into the generalized born model. J Chem Phys 2005; 122:094511.
85. Simonson T. Free energy calculations: approximate methods for biological macromolecules. In Chipot C, Pohorille A, editors. Free energy calculations: theory and applications in chemistry and biology. New York: Springer Verlag; 2006, Ch. 12.
86. Gallicchio E, Kubo M, Levy R. Enthalpy-entropy and cavity decomposition of alkane hydration free energies: numerical results and implications for theories of hydrophobic hydration. J Phys Chem B 2000;104:6271–6285.
87. Archontis G, Simonson T. Proton binding to proteins: a free energy component analysis using a dielectric continuum model. Biophys J 2005;88:3888–3904.
88. Pace C, Scholtz J. A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 1998;75:422–427.
89. Guex N, Peitsch M. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.

3.2 Ajustement de la fonction d'énergie et évaluation du modèle CASA pour différentes applications liées au CPD

Dans une seconde phase, nous avons poursuivi l'évaluation du modèle CASA dans deux directions. (1) Nous avons examiné si le modèle s'améliore avec un paramètre spécifique pour les aromatiques (atomes impliqués dans les groupements aromatiques). (2) Nous avons examiné deux nouveaux problèmes : interaction protéine-ligand et *protein design*. Par ailleurs, nous avons aussi mis en place un modèle amélioré pour l'état déplié.

3.2.1 Réoptimisation des coefficients de solvatation atomiques

Un quatrième paramètre de solvatation pour les atomes impliqués dans les groupements aromatiques fut ajouté à notre précédent jeu de paramètres 'Fraternali Modifié' ou 'MF'. Nous avons alors complètement réoptimisé les quatre paramètres de solvatation, associés respectivement aux groupements neutres, aromatiques, polaires et ionisés. Un jeu de 140 mutations expérimentales fut utilisé pour cette reparamétrisation. Le jeu de paramètres résultant de cette optimisation sera appelé le jeu PHIA, pour 'Polar Hydrophobic Ionic Aromatic'.

D'autre part, la prédiction de changements de stabilité requiert une bonne description de l'état déplié. En effet, la stabilité d'une protéine est déterminée par la compétition entre sa structure native et un ensemble de structures dépliées. Nous avons donc mis en place un modèle légèrement amélioré pour l'état déplié. Nous sommes partis du modèle de tripeptide ala-X-ala, décrit en section 1.1.2. La chaîne latérale de chaque acide aminé n'interagit qu'avec le squelette environnant et le solvant. Pour chaque type d'acide aminé X, fut considéré un grand nombre de structures de tripeptides possibles avec différentes conformations du squelette et de la chaîne

latérale. Ces différentes structures furent extraites de six protéines de structures connues. Pour chaque acide aminé X, la structure conduisant à la meilleure énergie fut identifiée et considérée comme la structure de référence. La détermination de ces structures de référence a dû se faire en même temps que l'optimisation du modèle de solvant PHIA. Pour ce faire, une procédure itérative fut mise en place. Dans un premier temps, les structures de références furent déterminées avec le jeu courant des paramètres de solvatation, puis les paramètres de solvatation furent à leur tour optimisés pour le jeu de structures de référence précédemment déterminées. Ces opérations furent répétées jusqu'à convergence. Après avoir optimisé notre fonction d'énergie pour la prédiction de stabilité, nous avons testé nos deux modèles CASA (MF et PHIA) dans diverses applications.

3.2.2 Evaluation du modèle CASA dans différentes applications

Tout d'abord, les différents modèles de solvant furent de nouveaux évalués pour la prédiction de changements de stabilité associés à des mutations ponctuelles. Cette fois, les mutations impliquaient tous les types d'acides aminés (l'étude précédente se limitait aux mutations impliquant des résidus chargés [134]). Nous avons ensuite testé la capacité de ces modèles dans la prédiction de changements d'affinité protéine-ligand ou protéine-protéine associés à des mutations. Cette application jouera un rôle très important dans l'évolution dirigée des synthétases. Enfin, nous avons réalisé le *design* complet de 8 domaines SH3. Ce travail est décrit dans l'article ci-dessous :

M. Schmidt am Busch, A. Lopes, N. Amara, C. Bathelt and T. Simonson. Testing the Coulomb/Accessible Surface Area solvent model for protein stability, ligand binding, and protein design, 2008. *BMC Bioinformatics, 9 :148.*

BMC Bioinformatics

BioMed Central

Research article

Open Access

Testing the Coulomb/Accessible Surface Area solvent model for protein stability, ligand binding, and protein design

Marcel Schmidt am Busch, Anne Lopes, Najette Amara, Christine Bathelt and Thomas Simonson*

Address: Laboratoire de Biochimie (UMR CNRS 7654), Department of Biology, Ecole Polytechnique, 91128, Palaiseau, France

Email: Marcel Schmidt am Busch - marcel.schmidt-am-busch@polytechnique.edu; Anne Lopes - anne.lopes@polytechnique.edu; Najette Amara - najette.amara@polytechnique.edu; Christine Bathelt - christine.bathelt@polytechnique.edu; Thomas Simonson* - thomas.simonson@polytechnique.fr

* Corresponding author

Published: 13 March 2008

BMC Bioinformatics 2008, **9**:148 doi:10.1186/1471-2105-9-148

Received: 9 October 2007
Accepted: 13 March 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/148

Abstract

Background: Protein structure prediction and computational protein design require efficient yet sufficiently accurate descriptions of aqueous solvent. We continue to evaluate the performance of the Coulomb/Accessible Surface Area (CASA) implicit solvent model, in combination with the Charmm19 molecular mechanics force field. We test a set of model parameters optimized earlier, and we also carry out a new optimization in this work, using as a target a set of experimental stability changes for single point mutations of various proteins and peptides. The optimization procedure is general, and could be used with other force fields. The computation of stability changes requires a model for the unfolded state of the protein. In our approach, this state is represented by tripeptide structures of the sequence Ala-X-Ala for each amino acid type X. We followed an iterative optimization scheme which, at each cycle, optimizes the solvation parameters and a set of tripeptide structures for the unfolded state. This protocol uses a set of 140 experimental stability mutations and a large set of tripeptide conformations to find the best tripeptide structures and solvation parameters.

Results: Using the optimized parameters, we obtain a mean unsigned error of 2.28 kcal/mol for the stability mutations. The performance of the CASA model is assessed by two further applications: (i) calculation of protein-ligand binding affinities and (ii) computational protein design. For these two applications, the previous parameters and the ones optimized here give a similar performance. For ligand binding, we obtain reasonable agreement with a set of 55 experimental mutation data, with a mean unsigned error of 1.76 kcal/mol with the new parameters and 1.47 kcal/mol with the earlier ones. We show that the optimized CASA model is not inferior to the Generalized Born/Surface Area (GB/SA) model for the prediction of these binding affinities. Likewise, the new parameters perform well for the design of 8 SH3 domain proteins where an average of 32.8% sequence identity relative to the native sequences was achieved. Further, it was shown that the computed sequences have the character of naturally-occuring homologues of the native sequences.

Conclusion: Overall, the two CASA variants explored here perform very well for a wide variety of applications. Both variants provide an efficient solvent treatment for the computational engineering of ligands and proteins.

Background

Solvation effects play an important role in protein folding and stability. Likewise, the processes of protein:protein and protein:ligand binding are accompanied by effects such as desolvation and rearrangement of solvent molecules. These solvation effects can be calculated by explicit solvent models, such as molecular dynamics simulations using a sphere of water molecules [1]. For certain large-scale applications, however, this explicit solvent treatment is too time-consuming. In protein design, an enormous number of amino acid sequences need to be considered [2]. Likewise, the screening of a large library of ligand molecules as a function of their binding affinity to a protein is a costly procedure and requires efficient methods. To avoid these problems, implicit solvent models are often used, which yield significant computational efficiency [3]. They do not consider the solvent degrees of freedom explicitly, but treat the solvent as a continuous medium having the average properties of the real solvent. Empirical methods, such as the solvent accessible surface area (ASA) model [4], often provide simple and quick ways of evaluating the solvation energy with an accuracy comparable to theoretical models. ASA models have become widely accepted within available implicit solvent treatments and have been used successfully in many applications, such as protein molecular dynamics [5-7], structure prediction [8] and protein:ligand binding [9,10]. In the ASA approach, the solvation free energy of a solute is expressed as a sum of atomic contributions, weighted by their solvent-exposed area. The contribution of each atom is quantified by a surface coefficient, which reflects the hydrophobicity or hydrophilicity of the particular atom type.

Apart from non-polar contributions to the solvation free energy, such as the entropy cost for cavity formation and van der Waals interactions, electrostatic contributions play an important role. Due to the fitting of the ASA model to experimental data, the electrostatic contribution is partly incorporated into the parameters. However, especially when using a small number of atom types, it is necessary to additionally calculate a screening energy, which accounts for the shielding of protein-protein electrostatic interactions by the high dielectric solvent. A simple approach is to add a term that reduces the electrostatic interactions between protein atoms by a constant factor, ε, which plays the role of a dielectric constant. To balance the two components of the solvent model, an overall weight α is applied to the ASA term. This combined model is known as the Coulomb/Accessible Surface Area (CASA) model [3]. More accurate approaches for screening energy calculations are the Generalized Born (GB) model [11,12] and the Poisson-Boltzmann (PB) model [13-15]. A disadvantage of these methods, however, is that they are not easily pairwise decomposable [16]: the energy is not expressed as a sum over atom pairs. Further, they are relatively time-consuming compared with surface area based models.

The first set of atomic solvation parameters distinguishing between different atom types was developed by Eisenberg and McClachlan in 1986 [4]. They used octanol to water transfer energies for 20 amino acids to derive solvation coefficients for 5 atom types. Subsequently, a number of studies have been devoted to the parameterization of the atomic surface coefficients. They differ in the assignment of atoms to characteristic groups and in the experimental data that were used to fit the coefficients. Ooi et al. [17] used 7 different atomic coefficients, fitted to experimental free energies of solvation of small organic molecules. Fraternali and van Gunsteren [6] restricted the atom types to only two: one for carbon, representing the hydrophobic effect, and one for both nitrogen and oxygen, representing the hydrophilic effect. These two parameters were optimized, such that the hydrophobic and hydrophilic solvent-accessible surface areas in proteins obtained from MD simulations matched those measured from the corresponding X-ray structures. Later, several more highly-parameterized ASA models [10,18,19] were developed, which use up to 100 different atom types and large training sets of experimental solvation free energies for diverse organic molecules.

In previous work [20], we optimized the CASA model for side chain placement and mutagenesis. We modified the atomic solvation parameters of Fraternali [6] by including an additional surface coefficient for atoms in charged groups, and by optimizing the dielectric constant ε and the weight α of the surface term. The model was optimized and tested using sidechain reconstruction calculations, protein solvation energies, and stability changes for point mutations involving the insertion or removal of a charged sidechain. In this paper, we continue to explore the performance of the CASA model, pursuing two directions. First, we consider whether a specific treatment of aromatic groups can lead to an improved parameterization. Second, we consider a broader set of test calculations than before. We again consider stability changes, but we include a wider variety of mutations, most of which do not affect charged groups. We consider the calculation of protein:ligand binding free energy changes due to point mutations, a very important application. So far, most ligand binding studies using ASA models used a large set of atomic surface coefficients. Pei et al [10], for example, used 100 atom types and reproduced the binding free energies of a test set of 50 protein:ligand complexes with a standard error of 2.0 kcal/mol. Our own model is much less heavily parameterized, but yields a comparable accuracy.

Finally, we perform automated protein design for eight small proteins of the SH3 family. Computational protein design is another area that requires good implicit solvent models [2,21,22]. This approach can help engineer new proteins as well as predict protein structure. It considers a given backbone structure of a protein and predicts the amino acid sequences that fold into it [23-30]. This information can be used either to identify mutations that stabilize a given structure or to assign the given 3D structure to new sequences with yet undetermined protein structures. The protein design procedure is applied here to a small test set of eight SH3 proteins and the performance of the optimized solvation parameters is compared to that of the earlier parameter set.

The newer CASA parameterization derived here treats aromatic groups as a specific group. All the model parameters are then reoptimized from scratch. Most studies so far have used experimental transfer free energies from octanol or cyclohexane to water for small model molecules to derive solvation parameters. However, organic solvents are only a crude approximation of the protein interior [9]. A more recent study by Zhou et al. [31] used a database of protein mutation experiments to develop atomic solvation parameters. With these parameters, the binding free energies of 21 protein-protein complexes were predicted with an rms deviation of 2.3 kcal/mol. Likewise, Lomize et al. [32] achieved very good agreement with experimental data using atomic solvation parameters based on protein stabilities. Here, we use a similar approach and derive our newer solvation parameters from experimental protein and peptide stability changes. We employ a procedure that attempts to match the computed stability changes to the experimental values, using a set of 140 mutations. Simultaneously, the model for the unfolded reference state of a protein is optimized: for each amino acid type, a preferred unfolded conformation is chosen from a large library of tripeptide fragments obtained from various proteins.

The final parameter set yields improved performance for protein stabilities, as expected. The newer and the earlier parameters yield comparable, good performance for ligand binding and protein design. Given the importance of implicit solvent models in the fields of structure-based drug design, prediction of protein structure and protein design, both of our optimized CASA models should be valuable tools for a wide range of applications.

Results

CASA parameter optimization and stability calculations

In recent work, we optimized and tested the CASA model, using just three atomic categories and performing sidechain reconstruction and protein stability calculations. We refer to the corresponding set of atomic surface coefficients as the "modified Fraternali", or MF parameters (Table 1). Indeed, the starting point for the earlier optimization was the parameter set of Fraternali & van Gunsteren [6]. Here, we want to test the CASA model further, and to explore whether a more extensive parameterization will improve its performance. Our previous CASA model [20] included only polar, unpolar and ionic atom coefficients. The current protocol also distinguishes aromatic atoms as a separate group. This refined atom assignment may be expected to improve the solvent model, as it has been shown that aromatic groups have solvation properties different from aliphatic groups [33]. With this modification, we reoptimized all the model parameters from scratch. We scanned a range of values for the four different surface coefficients; additionally, three different values for the dielectric constant ε in the Coulomb screening term (Eq. 2) were tested. A data set of 140 experimental stability mutations was used to adjust the parameters: these data correspond to 7 different helical peptides and 3 different proteins. The protein mutations involve mainly ionized residues, while the peptide mutations (which were taken from helix propensity scales) cover the full range of amino acids. In contrast to an earlier study by Lomize et al. [32], which derived atomic solvation parameters based on stability data for the protein interior, and in contrast to our previous study [20], this set of mutations involves mainly solvent-exposed residues (14% buried residues).

To compute the stability of a protein, it is necessary to construct a model for the unfolded state. Here, we assume that sidechains do not interact with each other in the unfolded state, and we describe the environment of each sidechain with a simple tripeptide model. Each sidechain thus interacts only with the local tripeptide backbone and the solvent. For each amino acid type, one preferred structure was chosen from a large set of tripeptide conformations obtained from various proteins.

Since the choice of the preferred unfolded reference structure for each amino acid type also depends on the parameterization of the solvent model, the parameter optimization had to proceed iteratively (Figure 1): (i) A preferred tripeptide structure was chosen for each amino acid using the current set of solvation parameters (ii); The solvation parameters were adjusted to best reproduce the experimental stability changes using the current set of reference structures. This was done based on the rms deviation of the computed stability changes from the experimental values.

Starting from the previously optimized solvation parameters [6,20], the optimization procedure converged after 4 iteration cycles. Table 1 shows the set of surface coefficients chosen from the top-ranking results. We refer to the

Table 1: Atomic solvation parameters (kcal/mol/Å²) for different atom types

atom type	MF	PHIA
unpolar	0.0119	-0.005
aromatic	0.0119	-0.04
polar	-0.0597	-0.08
ionic	-0.15	-0.10

MF: modified Fraternali parameters; PHIA: parameters optimized here.

Table 2: Rms and mean unsigned error (kcal/mol) for stability mutations

group of data	number of mutations	rms error	mean error
all	140	2.94	2.28
peptides	67	2.68	2.20
proteins	73	3.17	2.34
charged	87	3.06	2.29
uncharged	54	2.72	2.22

new set as the PHIA set (for "polar, hydrophobic, ionic, aromatic"). Compared with the earlier, MF parameters, the PHIA set shows a clear difference between the aromatic and the unpolar coefficient, a more positive ionic coefficient and a slightly more negative unpolar coefficient. The experimental and computed stability changes are compared in Figure 2. The correlation between the two sets is modest, 43%. Nevertheless, the mean error is reasonable, with an rms deviation of 2.94 kcal/mol for the complete set of experimental values and a mean unsigned error of 2.28 kcal/mol (Table 2). The rank ordering of the data is characterized by a Spearman rank correlation of 29% [34]; the probability of obtaining this value by chance is less than 0.001 (according to Student's test with a *t*-value of 3.5 and 140 degrees of freedom [34]). The PHIA results represent a significant improvement over the performance of the MF parameters, which give an rms deviation of 4.65 kcal/mol and a mean unsigned error of 3.56 kcal/mol. The dielectric constant for the new, PHIA set of surface coefficients was set to 24. A higher dielectric constant of 32 gives only insignificant improvement (rms = 2.93 kcal/mol and mean = 2.27 kcal/mol), while a lower value of 16 leads to noticeably poorer agreement with experiment (rms = 3.23 kcal/mol and mean = 2.52 kcal/mol).

In addition to the criterium of minimal deviation from the experimental data, care was taken to select a parameter set that makes sense physically. The ordering of the coeffi-

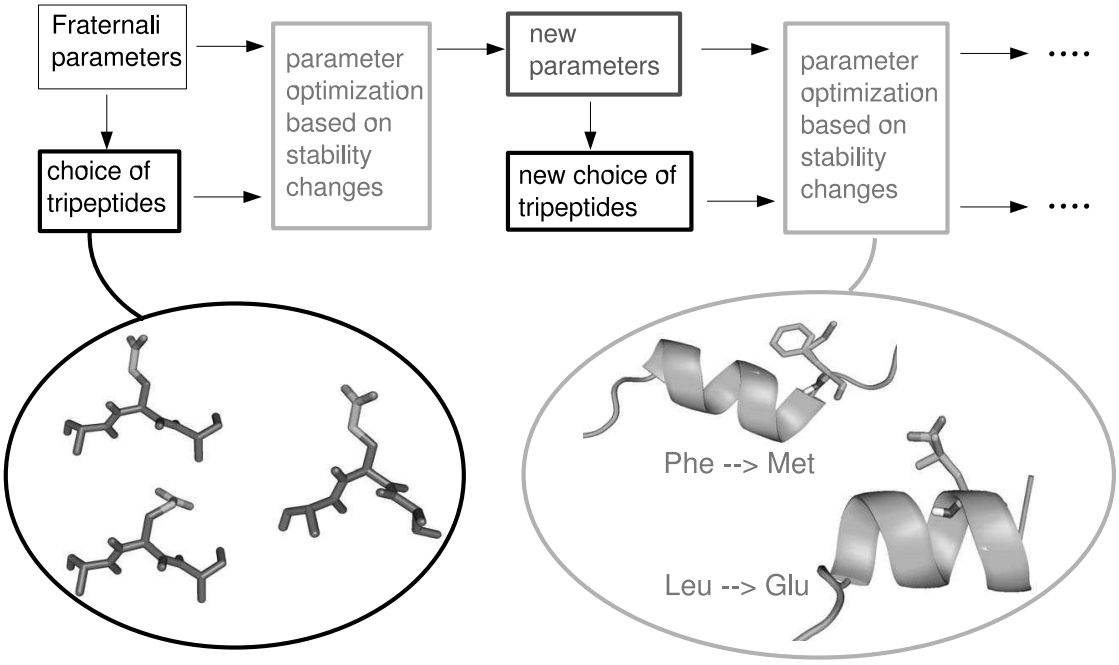


Figure 1
Iterative parameter optimization. Iterative optimization of the atomic solvation parameters and the tripeptide models for the unfolded reference state of a protein.

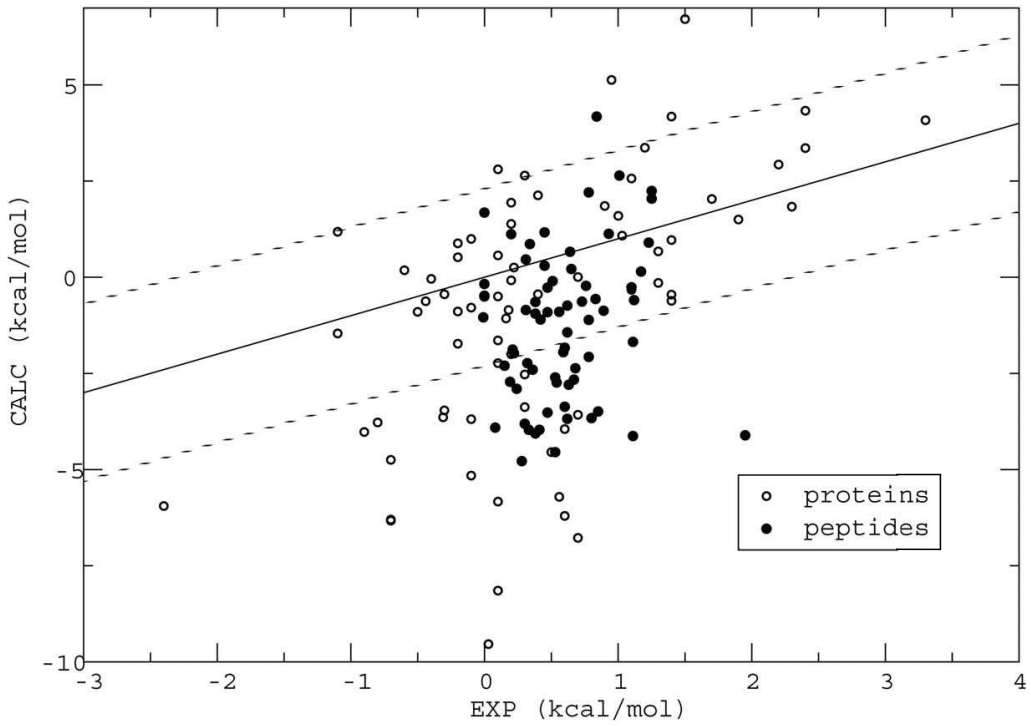


Figure 2
Stability changes. Calculated and experimental changes in stability upon mutation for 7 helical peptides and 3 proteins. The solid line corresponds to a (rather poor) linear fit; it and the dashed lines should be viewed as simple guides to help appreciate the error magnitudes.

cients should follow the expected preference of solvent exposure of each atom group: unpolar < aromatic < polar < ionized. Further, the coefficients of different groups should be sufficiently separated from each other. The surface coefficient of the unpolar atom group was allowed to be slightly negative, as this led to better agreement with the experimental data. To ensure that this does not lead to unphysical behaviour, the surface coefficient was tested on a propane dimer system. Studies on methane and neopropane association in water report an association free energy of -1.0 and -2.7 kcal/mol, respectively [35,36]. Using values between 0.0119 and -0.01 kcal/mol/Å^2 for the nonpolar surface coefficient, the computed association free energy varies from -2.75 to +0.09 kcal/mol. The nonpolar coefficient of -0.005 kcal/mol/Å^2 used here gives an acceptable association free energy of -0.56 kcal/mol.

To verify that the parameters were not overly biased by the optimization procedure, we performed cross-validation tests (see Methods). Part of the data (30 mutations) were omitted from the optimization process, then used to test the error level. This led to very similar parameters and errors compared to the original parameter optimization. Specifically, one cross-validation run led to the following optimized parameters (in kcal/mol/Å^2; the PHIA values obtained above are given in parentheses): aromatic: -0.08 (-0.04); ionic: -0.10 (-0.10); polar: -0.09 (-0.08); nonpolar: -0.005 (-0.005). The optimal dielectric was 24, as before. The mean and rms errors for the omitted data were 2.22 and 3.13 kcal/mol, respectively, compared to 2.11 and 2.70 for the optimization data. The mean and rms errors with the PHIA parameters for the omitted data were similar: 2.32 and 2.92 kcal/mol. The other cross-validation run led to the exact same atomic coefficients and dielectric constant, and to mean and rms errors for the omitted data of 2.08 and 2.68 kcal/mol (compared to 2.16 and 2.71 kcal/mol with the PHIA set). Thus the optimized parameters and the error levels are similar with and without cross-validation, showing that the resulting parameters are fairly robust.

The model performance shows a moderate dependence on the system considered. The peptide models taken alone lead to a slightly better agreement with the experimental data (Table 2), which might be due to their simple helical structure compared with a large protein system

(see also Figure 2). The performance of the uncharged mutations is better than that of the charged ones (Table 2). This occurs partly because the proportion of charged mutations is higher for the proteins (mostly charged mutations).

Compared with the results of Lomize et al. [32], which gave an rms deviation from experimental stability changes of only 0.41 kcal/mol, the performance here is less good. In their study, however, the solvation parameters were fitted to a more restricted set of experimental data. Only buried, uncharged residues in α-helices and β-sheets of proteins were used. In contrast, in our study, both charged and uncharged, solvent-exposed and buried, and protein and peptide mutations were included. Therefore, it can be expected that our more varied set of mutations results in a higher deviation from experiment.

Binding affinities

The performance of both the earlier, MF and the new, PHIA solvation parameters was assessed on a large set of experimental binding affinities and resistance mutations. These data include 80 mutations in six different ligand-protein systems: (i) tyrosyl-tRNA synthetase in complex with tyrosine (TyrRS), (ii) aspartyl-tRNA synthetase in complex with aspartate (AspRS), (iii) lysozyme in complex with the antibody HyHel-10 (Lyso) and (iv) the complex of the transmembrane glycoprotein CD4 with the gp120 component of the HIV virus (CD4); (v) the BPTI:trypsin complex; (vi) the chymotrypsin:BPTI complex. A seventh system was also studied: the tyrosine kinase Abl in complex with the drug imatinib. For this system, information is only available on mutations leading to resistance to the drug imatinib [37], while for the other four systems, more precise values of the binding affinities are available. The energy function was slightly different from the one used for the stability mutations above. Here, the dielectric constant (Eq. 2) ε was set to 16 and the weight α of the surface area term was set to 0.5 (instead of ε = 24, α = 1, above). These values gave improved performance for the binding affinities.

For the first four systems, (i-v; 55 mutations in all), the calculation of binding affinities followed a very simple protocol that only optimizes the conformation of the side chain in the mutated position, while the rotamers of the remaining side chains are left unchanged apart from a slight minimization. Further, the starting structures of the ligand-bound state and the ligand-free state are assumed to be identical except for the side chain at the mutated position. Although this is a rather simple approach, the resulting computed binding affinities are in reasonable agreement with the experimental values. The mean unsigned error for the differences in binding affinities upon mutation is 1.76 kcal/mol with the PHIA parameters, and 1.47 kcal/mol with the MF parameters (Table 3). Thus, the performance of the two CASA variants for this application is similar. The PHIA results for all mutations are given in Additional File 1. The performance is slightly poorer when charged residues are mutated. The more complicated systems such as protein-antibody binding (Lysozyme) or protein-protein binding (CD4) give slightly higher deviations from experiment (see also Figure 3). The mean unsigned errors for the small molecule-ligands (AspRS and TyrRS), combined, is 1.24 kcal/mol with PHIA, while the protein-ligand systems (Lysozyme and CD4), combined, give a mean unsigned error of 2.29 kcal/mol (Table 3).

Results were also computed for two other systems, the BPTI:trypsin [38] and BPTI:chymotrypsin [39] complexes. 13 mutations at position 15 in BPTI were studied in each case. The native residue is a lysine. One mutation was excluded (K15W in BPTI:trypsin) due to a large van der Waals contribution to the affinity change (see Methods), leaving 25 mutations. With the simple protocol used above, the agreement with experiment was poorer, with a mean error of 3.4 kcal/mol for the 25 mutations. A slightly different protocol was then tried. The entire BPTI

Table 3: Mean error (kcal/mol) for the binding free energies with CASA and GB/SA

data set	number of mutations[b]	CASA	GB-ACE	GB-HCT	reference
all[a]	55 (52/48)	1.76	2.39	1.96	[66–72]
charged[a]	24 (23/20)	2.09	2.70	2.89	
AspRS	9 (9/9)	1.86	3.40	1.25	[73]
TyrRS	15 (15/15)	0.62	0.80	1.13	[66–70]
Lyso	9 (9/7)	2.68	3.01	2.92	[71]
CD4	22 (19/17)	1.89	2.34	2.53	[72]
BPTI[c]	25	2.75	-	-	[38, 39]

[a]Excluding the BPTI complexes, which were computed with a slightly different protocol and for which GB data are not available. [b]For GB-ACE and GB-HCT, some mutations were excluded due to unfavorable VdW contacts; the numbers of mutations (ACE/HCT) are in parentheses.
[c]BPTI:trypsin and :chymotrypsin complexes.

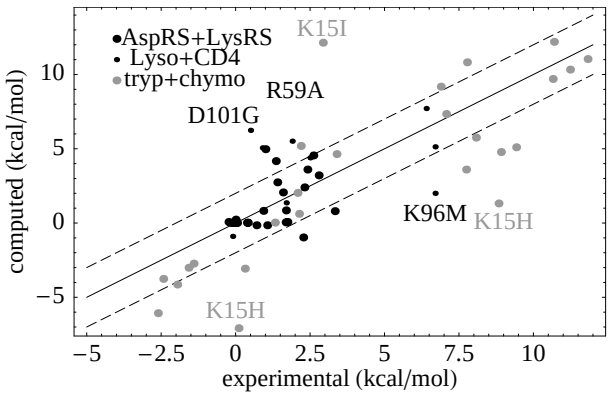


Figure 3
Binding affinities. Calculated and experimental differences in binding affinity upon mutation for 6 different protein-ligand systems: Aspartyl- and Tyrosyl-tRNA synthetases (AspRS+TyrRS), a Lysozyme-antibody complex (Lyso), the CD4 complex with the gp120 component (CD4), and the BPTI complexes with trypsin and chymotrypsin.

was minimized, instead of just the mutated position (see Methods). This led to improved agreement, with mean errors of 2.68 and 2.81 kcal/mol for BPTI:trypsin and BPTI:chymotrypsin, respectively: about the same level as for the lysozyme:antibody complex. This illustrates the need for more extensive structural relaxation with some mutations. More generally, these two examples show that, not surprisingly, with the simple energy functions and conformational exploration used here, a certain amount of system-specific parameter fitting and adjustment can be necessary.

For the resistance mutations in Abl-kinase, good qualitative agreement was achieved, in so far as all the mutated proteins gave less favorable binding affinities than the native protein (Table 4). For positions 22 and 86, the computed difference between the mutant and native binding free energies may be too small; the other three values, however, are clearly consistent with the observed resistance to imatinib binding.

Table 4: Binding free energy differences for the ABL:imatinib complex with CASA and GB/SA

mutation	CASA	GB-ACE	GB-HCT
L17V	1.74	1.00	0.80
Y22F	0.30	0.23	0.10
V58A	6.50	1.23	1.10
F86L	0.46	0.75	2.50
F128V	5.99	5.46	7.10

In kcal/mol.

There are several mutations that are badly predicted, including R59A in the CD4:gp120 system and D101G and K96M in the Lysozyme:antibody system (Figure 3). The first two cases involve a large to small mutation, and all three involve removal of a net charge; these processes which might require a more extensive rearrangement of the surrounding residues, as in the BPTI complexes (where the native residue was a lysine in all 25 mutations). In the simple protocol employed here, the rotamers of the side chains in the vicinity of the mutation are not reoptimized, and the slight minimization carried out here might not be sufficient to model a realistic mutated protein conformation. For the K96M case, we tried the same protocol as for the BPTI cases, applying 50 steps of Powell minimization to the entire lysozyme protein, instead of just the mutated sidechain. This led to a computed binding free energy change of +4.5 kcal/mol, much closer to the experimental value of 7.9 kcal/mol. For one other case, the D78A mutation in TyrRS, we employed a much more extensive conformational search: all the rotamers close to the mutation site were explored using a stochastic search strategy; the error in the binding free energy decreased from 1.9 kcal/mol to 0.8 kcal/mol (not shown). These examples suggest that when the sidechain charge and/or volume changes substantially, a more sophisticated conformational sampling is required. More work in this direction is underway.

On the whole, however, we obtain fair agreement with experimental data using a very simple method. The overall mean unsigned error for the 80 mutations is 1.96 kcal/mol; the rms error is 2.73 kcal/mol. The correlation between the computed and measured data is 73%. The rank ordering of the data is characterized by a Spearman rank correlation of 70% [34]; the probability of obtaining this value by chance is less than 0.001 (according to Student's test with a *t*-value of 8.59 and 80 degrees of freedom [34]). Our errors are only slightly higher than those reported by Pei et al., who obtained an rms error of 2.0 kcal/mol for the binding affinities of 50 protein-ligand complexes. Likewise, the accuracy obtained here appears comparable to that achieved by Guérois et al. [40] for their charged mutations, although this subset was not reported separately. Recently, Handel et al. [41] predicted the stabilities of more than 1500 mutants to within 1 kcal/mol. Partly, this improvement might be due to their more sophisticated, all-atom, force field, which increases accuracy but makes their calculations almost an order of magnitude more expensive, compared with the CHARMM19/CASA level (the increased cost being due partly to the explicit treatment of all hydrogens and partly to the need for a more detailed rotamer library). Further, our data set contains a higher percentage of charged mutations, which makes predictions more demanding.

As an additional reference, the performance of the optimized CASA models was compared to that of a GB/SA model, which should provide a more accurate treatment of the electrostatic contribution to the solvation free energy [42]. Only the 55 mutations for systems (i-v) were studied. Two variants of the GB model were tested: (i) GB-HCT [43] in combination with the Amber, all-atom force field [44] and (ii) GB-ACE [45,46] with the CHARMM19 force field [47]. For the GB-HCT variant, the overall quality of the results is close to the CASA level, with a slightly higher mean error (1.96 kcal/mol, Table 3). The GB-ACE variant gives notably higher mean errors for all protein-ligand systems. This might be due to the GB-ACE parameterization which is not optimized for this specific application. The GB-HCT parameters were optimized previously for computational sidechain placement and protein mutagenesis [20]. Some differences between GB-HCT and CASA might be due to the different force field treatments, as Amber uses explicit hydrogens on all atoms, while CHARMM19 uses implicit hydrogens for unpolar atoms. The qualitative agreement with the experimental resistance mutations for Abl:imatinib for the two GB variants is comparable to that obtained using the optimized CASA models (Table 4).

Protein design

In protein design, the amino acid side chains are mutated and sequences are selected to optimize the folding free energy, using a heuristic search algorithm. In our previous protein design study [48], we obtained good results for 16 different globular proteins using the MF solvation parameters. Here, we consider a subset of those proteins, consisting of 8 SH3 domains. These proteins are used to assess the performance of the new, PHIA solvation parameters for protein design. The protein design calculations were carried out using our Proteins@Home distributed computing platform, with the help of volunteers in several countries. Proteins@Home is discribed in more detail elsewhere [49].

For each protein, 450,000 sequences were generated and amino acid identities relative to the native sequence were calculated. Here, we give (consistent with the literature) the full identities of the designed sequences, even though Cys, Gly, and Pro were not allowed to mutate; see Tables 5 and 6. For further analysis, we selected two subsets of the computed sequences: (i) the 40 sequences with the highest identity scores relative to the native protein ("high-scoring sequences"); (ii) the 40 sequences with the best folding free energies ("low energy sequences"). For these two subsets, the mean sequence identities relative to the corresponding native sequence are given in Table 5. We will not discuss the mean identities over all 450,000 computed sequences, because the values lie very close to those obtained for the low energy sequences. For the low energy subset, the average sequence identity obtained with the PHIA parameters is 32.8%, slightly lower than the value obtained earlier with the MF parameters, 35.0% [48]. For the eight proteins, the identities obtained with PHIA range from 26.5% to 45.1%. The relative performance of the MF and PHIA parameters depends on the protein: since Hck and c-Src are better predicted with PHIA, while 1gcqB, Crk, Abl, and especially Csk are better predicted with MF. Considering the subset of high-scoring sequences, the PHIA parameters give results of roughly the same quality as MF: the mean identities for this subset are 44.4% with PHIA, compared to 46.6% with MF.

With both the PHIA and the MF parameters, the identity scores obtained for the 8 proteins lie within the range of published average identity scores for redesigned proteins [27,28,41,50,51]. In a protein design study by Jaramillo et al. [50] sequence optimizations for 11 SH3 domains were performed and resulted in an average sequence identity of 23.9%. Our energy-ranked PHIA sequences lie well above this score, with a sequence identity of 32.8% averaged over all 8 proteins. Recently, Saunders et al. [28] used a refined protein design method and reported sequence identities as high as 37% for 42 globular proteins. Considering the full sequence identities of our best-scoring

Table 5: Mean identities (%) for the computed sequences

			MF (ε = 10)		PHIA (ε = 14)	
PDB code	Name	length	low energy	high score	low energy	high scoring
1gcq(B)	Grb2	57	37.2	49.5	36.1	47.0
1gcq(C)	Vav	69	45.6	55.0	45.1	52.4
1cka	c-Crk	56	39.8	52.1	35.9	49.9
1shg	alpha-spec.	57	26.9	37.6	26.5	39.5
1abo	Abl kinase	58	37.7	48.5	34.7	45.8
1ad5	Hck kinase	58	22.6	38.4	23.6	36.1
1csk	Csk	56	37.3	48.5	27.0	38.5
1fmk	c-Src	60	32.4	43.5	33.8	45.9
	average		34.9	46.6	32.8	44.4

Table 6: Blosum scores for computed and natural sequences

PDB code	natural sequences	computed (all)	low energy	high scoring
1gcqB	98.2	67.8	74.0	100.8
1gcqC	43.2	110.5	121.6	141.2
1cka	99.1	57.8	71.2	104.1
1shg	95.9	35.7	38.5	65.2
1abo	89.0	67.3	71.4	98.3
1ad5	111.3	35.5	36.8	75.0
1csk	87.7	60.9	63.0	90.5
1fmk	122.2	56.2	64.2	92.5

sequences, both our parameter sets give results that lie close to this value. Pokala and Handel [41] used an all-atom force field and a GB/SA solvent model to redesign 8 proteins and achieved somewhat higher sequence identities, between 33.5 and 46.7%. This approach, however, includes a negative design criterion, which constrains the surface amino acid composition of the proteins to be native-like. It also leads to an increase in computational effort of about two orders of magnitude, compared with the CASA solvent model and a united-atom force field such as Charmm19.

Next, we addressed the question whether the sequences computed with the new solvation parameters ressemble naturally occuring sequences. As a reference, we created a set of natural sequences in the following way. For each of the 8 proteins, sequences from the SwissProt database with a sequence identity of more than 60% relative to the respective native sequence were retrieved. These sequences were then combined into a large set consisting of 94 sequences. Table 6 compares the average Blosum62 scores obtained for the natural sequence set with those for the complete set of computed sequences and for our two sets of high-ranking sequences (either low energy or high Blosum62 sequences). A weighting scheme was applied to the Blosum scores to account for the variability of each amino acid within the natural sequences. This scheme gives a greater weight to amino acids that are conserved within the natural set. For a frequency of an amino acid of more than 80% at a certain position within the natural sequence set, a weight of 1 was assigned; for a frequency between 50 and 80%, a weight of 0.75 was assigned; for a frequency of less than 50%, a weight of 0.5 was assigned. The Blosum62 scores show that the computed sequences cover a considerable part of the natural sequence set. Figure 4 shows that the computed scores overlap with the range of natural sequences. For most proteins, the weighted average Blosum score lies below the corrispinding score for the natural sequence set, but is within the range of the natural sequences. In one case (1gcqC), the weighted average Blosum score even exceeds the value of the natural set. This indicates that the computed sequences do indeed show native-like characteristics and behave like distant homologues of the native sequences.

Discussion

Simple, efficient, solvent models are of great importance in protein modelling and structural bioinformatics. Here, we have continued to explore the performance of the CASA solvent model, in two directions. First, we considered a variant of increased complexity, where aromatic atoms are treated as a separate group. Our previous approach [20] did not distinguish between unpolar and aromatic atoms. Indeed, the solvation properties of aromatic groups are rather different from other nonpolar groups found in proteins. For this variant, we reparameterized the model completely, leading to the PHIA parameter set. Second, we applied the CASA model to a wider set of applications than previously. We considered protein stability changes associated with point mutations, including all amino acid types (in contrast to our previous work [20]). We also considered protein:ligand binding, an especially important application. Finally, we performed complete protein redesign with the new parameters.

For the new, PHIA parameterization, four different atom types were considered for the atomic surface coefficients: unpolar, aromatic, polar and ionized atoms. The solvation parameters were fitted to a set of 140 experimental stability changes for protein and peptide mutations. Protein stabilities were calculated using an unfolded reference state modelled by a collection of tripeptide structures. These reference structures were taken from a large library of structural fragments from six different proteins. Starting from the earlier, MF solvation parameters [6,20], an iterative procedure was employed, which optimizes, at each cycle, both the solvation parameters and the model for the unfolded reference state of a protein. Atomic parameters were chosen that gave a minimal deviation from the experimental data and also represent the expected relative hydrophobicities of the atom groups. The selected parameter set gives a mean unsigned error of 2.28 kcal/mol for the 140 stability mutations, compared to 3.56 kcal/mol

with the MF parameters. Cross-validation tests gave similar parameter values and similar error levels.

For the binding calculations, over 50 experimental mutations in 5 different protein-ligand systems were used, including both small molecule ligands and protein-protein complexes. The calculated differences in binding free energy are in reasonable agreement with the experimental data, with both CASA variants. The mean unsigned error is 1.76 kcal/mol with the PHIA parameters and 1.47 kcal/mol with the MF parameters. It was also shown that the optimized CASA models are not inferior to methods such as GB-ACE/SA or GB-HCT/SA, which treat the electrostatic contribution to solvation more accurately. Two additional protein-protein complexes (BPTI:trypsin, BPTI:chymotrypsin) required a slightly modified protocol to give comparable error levels.

Protein design was carried out for eight SH3 domain proteins and the performance of the PHIA parameters was compared with that of the MF parameters. On average, slightly lower sequence identities with the native sequence were achieved using the new, PHIA parameters. Differences, however, are small, and the performance depends on the particular protein. On the whole, both parameter sets give sequences of comparable quality, which are competitive with recently published results for designed proteins. Further, the computed sequences were found to have the character of naturally occuring, distant homologues of the native sequences.

Conclusion

Overall, the CASA model performs well for a wide variety of applications. The PHIA parameters, specifically optimized here for protein stability, give a distinct improvement over the earlier, MF parameters for this application. For ligand binding and protein design, the specific treatment of aromatic groups in the PHIA parameterization did not lead to an improvement in performance. Rather, the two parameter sets perform well for both applications, with the exact relative performance dependening on the particular system. Both variants provide an efficient tool for the computational engineering of ligands and proteins.

Methods

Effective energy function

The effective free energy function we used in our calculations takes the following form:

$$E = E_{bonds} + E_{Angl} + E_{Dihe} + E_{impr} + E_{vdW} + E_{Coul} + E_{solv} \quad (1)$$

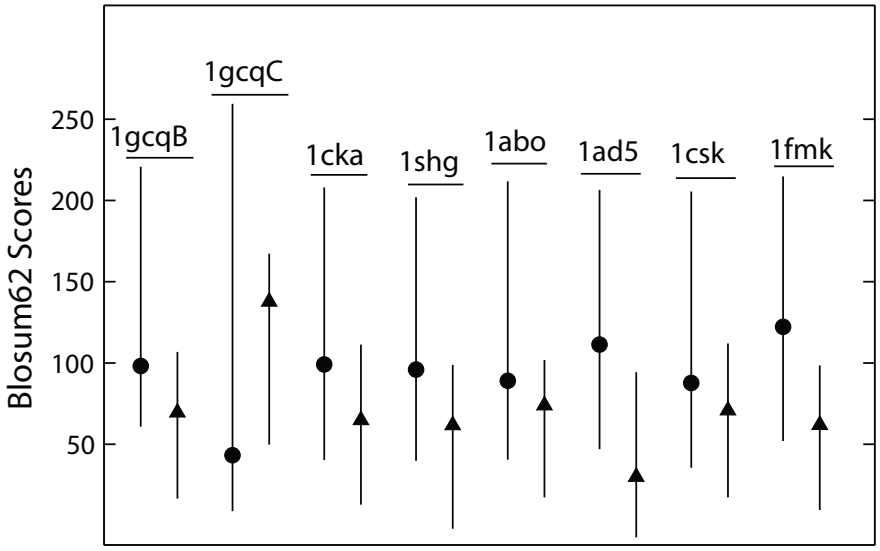


Figure 4
Comparison of Blosum scores for natural and computed sequences. For each protein (denoted by the respective PDB code), the vertical lines represent the range of scores within the natural sequence set (left) and the computed sequence set (right). The average score is shown as circles and as triangles for the natural and the computed sequences, respectively.

The first six terms represent the protein internal energy and are taken from the CHARMM19 empirical energy function [47]: a covalent bond energy term, a bond angle energy term, a torsion energy term, an improper dihedral energy term which maintains the chirality or planarity of certain atom centres, a Van der Waals energy term and a Coulomb electrostatic energy term. The last term, E_{solv}, models the effect of the solvent, and represent either a CASA term or a GB term in this study. When using the GB variant HCT for the solvent term, force field parameters for the energy function were taken from Amber [44] (see below).

Coulomb/Accessible Surface Area (CASA) model

This implicit solvent model uses a screened Coulomb energy term and a solvent accessible surface energy term [3]. The former describes the dielectric screening of solvent-solute interactions. It reduces interactions between solute atoms by a constant factor ε to account for the shielding by the high dielectric solvent. The latter term represents local solute-solvent interactions, such as van der Waals energy and cavity energy (creating a cavity against the solvent pressure and reorganization of solvent molecules around the solute), that are assumed to be proportional to the solvent accessible surface area of the solute atoms. The equation for the solvation free energy takes the following form:

$$E_{solv} = E_{screen} + E_{surf} = \left(\frac{1}{\varepsilon} - 1\right) E_{coul} + \alpha \sum_i \sigma_i A_i. \quad (2)$$

where A_i is the exposed solvent accessible surface area of atom i, and the summation is over all atoms in the solute; σ_i (measured in kcal/mol/Å^2) is a parameter that depends on the nature of atom i and reflects each atom's preference to be exposed or hidden from solvent; α is an overall weight applied to the surface energy term. Surface areas were computed by the Lee and Richards algorithm [52], implemented in the XPLOR program [53], using a 1.5 Å probe radius. The solute atoms were divided into 4 groups with characteristic surface coefficients σ_i: unpolar, aromatic, polar and ionic. Hydrogen atoms were assigned a surface coefficient of 0. The weight α was not optimized during the parameter scans (fixed to 1) but was adjusted in subsequent applications. For the protein design calculations (see below), a value of 1 was used. For the ligand-binding calculations (below), a value of 0.5 worked best. The dielectric constant was optimized in the parameter scans, with a value of 24 working well for the stability mutations. Values of 16 and 14 were used, respectively, for the ligand-binding and protein design calculations.

Generalized Born/Surface Area (GB/SA) model

A more sophisticated description of electrostatic interactions in a heterogeneous dielectric medium is provided by the Poisson-Boltzmann (PB) equation [13-15]. Given a spatial charge distribution in an environment with one or more dielectric constants, the electrostatic potential can be calculated. The cost involved in solving the PB equation numerically, however, limits its use. A more efficient alternative is the Generalized Born (GB) model [11,12], which is based on PB theory but replaces the solution to the electrostatic potential by an approximate calculation of the solvent-induced reaction field energy. In the GB/SA method, the electrostatic contribution to the solvation free energy is described by a GB term, while the non-polar contribution is modelled as proportional to the solvent accessible surface area of the solute. The two contributions are balanced by a factor σ applied to the surface area term:

$$E_{solv} = E_{GB} + \sigma \sum_i A_i \quad (3)$$

where E_{GB} is the GB term consisting of a self-energy term and an interaction term as described elsewhere [14,20]; A_i is the exposed solvent accessible surface area of atom i, and the summation is over all atoms in the solute. In effect, a single surface coefficient σ is used for all atom types.

Calculation of stability changes

In general, introducing a mutation changes the stability of a protein. Here, differences between the stability of a mutant protein and the native protein were calculated. The stability of each protein is computed as the difference in free energy between the folded state and an unfolded reference state. The free energy change upon mutating a protein is thus:

$$\Delta\Delta G = (G_{mut} - G_{mut}^{ref}) - (G_{nat} - G_{nat}^{ref}) \quad (4)$$

where G_{mut}, G_{nat} are the free energies of the mutant and native protein, respectively, and G_{mut}^{ref}, G_{nat}^{ref} are the free energies of the unfolded reference state for the mutant and native protein, respectively. The free energy of each state is evaluated using the effective energy function given in equation (1) with the solvent contribution being represented by a CASA term. The nonbonded interactions were cut off at a distance of 10 Å between atoms using a shifting and a switching function for electrostatic and van der Waals interactions, respectively.

The native structures for the folded state were taken from the Protein Data Bank (PDB) [54] with the structure codes

2LZM, 2RN2 and 1STN, respectively. The side chains were slightly minimized prior to any energy evaluation. For the peptide mutations, experimental structures are not available, and models were built using the SwissPDB viewer [55], which constructs an ideal α helix from a given sequence. After side chain minimization of the helix models, the structures were treated identically to the protein structures.

The corresponding mutant protein and peptide structures were created by replacing the side chain at the mutated position with the mutant side chain while maintaining all other atom coordinates. The coordinates for the mutant side chain were taken from the Tuffery rotamer library [56]. For each rotamer, the side chain was minimized (30 steps of Powell minimization), with the backbone fixed, and the energy of the mutant protein was evaluated. The mutant side chain rotamer giving the lowest energy for the mutant protein was retained. For the native structures, the side chain conformation at the position to be mutated was kept as in the crystal structure and only subjected to a short minimization (30 steps). In some cases, unfavourable van der Waals contacts in the proteins occured upon mutation. These data were considered as outliers and were not used for parameter adjustment (see below).

The experimental stability changes for the protein mutations were taken from the ProTherm database [57]. For the peptide stability changes, experimental helix propensity scales were used. The sequences of the various peptide systems are given below, with the mutated position denoted by X:

pepT1: SSDVSTAQXAAYKLHED [58],

KEAKE: YEAAAKEAXAKEAAAKA [59],

K2AE2: YSEEEEKAKKAXAEEAEKKKK [60],

VAR: KETAAAKFERQHMDS [61],

PAD: YKAAAAKAAXAKAAAAK [62],

KAL: YSEEEEKKKKXEEEEKKKK [63],

SH1: AETAAAKFERQHM [64],

SH2: KETAAAKFERAHA [64]

Unfolded state

In the unfolded state of a protein, it is assumed that amino acid side chains do not interact with each other, but only with nearby backbone groups and with solvent. This situation can be modelled by a collection of n tripeptide structures with the sequence Ala-X-Ala; n is the number of amino acids in the protein. For each amino acid type X, a number of possible structures with different backbone and side chain conformations were considered. These structures were extracted from various positions in the X-ray structures of 6 different proteins taken from the PDB [54]: lysozyme (2LZM), bovine pancreatic trypsin inhibitor (4PTI), staphylococcal nuclease (1STN), α-toxin (1PTX), ribonuclease A (2RN2) and cyclophilin (2CPL). In each tripeptide structure, the side chain X was slightly minimized with respect to itself and the backbone of the whole tripeptide. To choose the optimal tripeptide structure for each amino acid type, the interaction between the respective side chain and the tripeptide backbone served as a criterium. Thus, for each amino acid type X, the tripeptide structure giving the lowest interaction energy was taken to represent the preferred structure for X in the unfolded state. The total free energy of the unfolded state is obtained by summing the contributions, E_X, of the n individual amino acids of the protein. When comparing the folding free energies of two sequences, only sidechain – sidechain and sidechain – backbone interactions are taken into account. Interactions between different portions of the backbone cancel, both in the folded and the unfolded state, so that no important interactions are missed through the tripeptide unfolded model.

Iterative optimization

Since the choice of the reference structure for each amino acid depends on the set of parameters used in the CASA model, the solvent parameters and tripeptide structures had to be optimized iteratively (Figure 1). As a starting point for optimization, we took the surface parameters developed by Fraternali and van Gunsteren [6], supplemented by an additional surface coefficient for ionic atoms [20]. In this initial parameter set, aromatic atoms were assigned the same surface coefficient as unpolar atoms.

The atom groups were assigned as follows: (i) unpolar: all alkane carbons, the carbonyl carbons of the protein backbone, and S; (ii) aromatic: Trp, Phe and Tyr aromatic ring carbons and nitrogens; (iii) polar: N/O atoms not belonging to ionized groups, N-C-N group in the His ring; (iv) ionic: guanidinium group of Arg, carboxyl group of Asp/Glu, N-C-N group in the ring for protonated His.

Using this solvation model, a first choice of reference structures was made. Then, stability changes were calculated according to equation (3) using the current set of reference structures. These stability changes were calculated for all combinations of surface parameters within a range of values given below (Table 7). Thus, for each combination of parameters, a set of $\Delta\Delta G$ values was obtained; a mean unsigned error and an rms deviation from the experimental stability changes were determined. The rms

Table 7: Range of solvation parameters (kcal/mol/Å²) and dielectric values ε scanned during the iterative optimization

atom type	range	interval
unpolar	-0.005 to 0.01	0.005
aromatic	-0.08, -0.06 to 0.01	0.01
polar	-0.12, -0.10 to -0.04	0.01
ionic	-0.20, -0.18 to -0.10	0.01
ε	16 to 32	8

deviation was used as a score for the performance of a given surface parameter combination. From the top-scoring parameter sets, a physically meaningful set was chosen and used as a new starting point for the next iterative optimization cycle. This procedure was repeated until no further significant change in the ranking of the parameter combinations occurred.

Cross validation

The optimization procedure was tested for bias and overfitting by the following cross-validation procedure. 30 of the 140 mutants were chosen randomly and left out of the optimization. The reference structures were taken from the above, iterative optimization and kept fixed. Parameter scanning was then performed, leading to several good quality parameter sets. The mean errors were then computed for the omitted, or "test" data. This procedure was done twice, with two distinct sets of mutations omitted from the optimization. The parameter sets and error levels from these two runs were similar to each other and to the iterative optimization described above, showing that the optimization is not subject to excessive overfitting or bias.

Binding affinities

Binding affinities were calculated for 5 different ligand-protein systems taken from the PDB [54] and a number of their mutants: (i) tyrosine kinase Abl in complex with imatinib (1OPJ), (ii) Tyrosyl-tRNA synthetase in complex with tyrosine (4TS1), (iii) Aspartyl-tRNA synthetase in complex with aspartate (1IL2), (iv) Lysozyme in complex with the antibody HyHel-10 (3HFM) and (v) the complex of the glycoprotein CD4 with the gp120 component of the HIV virus (1G9M). As starting structures, the ligand-bound X-ray structures of these 5 proteins were used. The system (iii) was truncated to a 30Å sphere around the ligand; systems (ii), (iv) and (v) were truncated to 20 Å spheres around the ligand, and for system (i) the untruncated structure of chain B was used. The mutant structures were created by replacing the side chain at the relevant position with a rotamer of the mutant side chain from the Tuffery library [56].

A simple protocol was adopted for the energy evaluation: The side chain at the mutated position was subjected to 50 steps of minimization with respect to itself and to all other side chains with the backbone kept fixed. During this minimization, all sidechains were allowed to adjust to the introduced mutation but otherwise inter-sidechain interactions were excluded. The energy of this slightly adjusted protein conformation was taken as the ligand-bound energy. For the ligand-free state, the ligand was removed, the side chains were again minimized and the energy of this conformation was taken. For the mutated sidechain, all rotamers from the library were considered, and the lowest energy for the ligand-bound and ligand-free states, respectively, was retained. In the native structure, the rotamer at the position to be mutated was not varied, as it is assumed that the X-ray structure already represents a low energy conformation.

Three different solvent treatments were employed in this protocol:

(1) The CASA model as described above using the parameters optimized earlier [20], with a weight factor α of 0.5 and a dielectric constant ε of 16. We obtained good results for sidechain placement and stability changes in our previous work [20] using these values.

(2) The CASA model with the parameters optimized here, with the same weight factor α of 0.5 and dielectric constant ε of 16. These values of α and ε were chosen because they gave the best agreement with the experimental binding affinities.

(3) A Generalized Born/Surface Area (GB/SA) model with a weight factor σ of -0.05 kcal/mol/Å^2 for the surface term and a dielectric constant ε of 8.0.

Mutations for which the van der Waals energy contributed more than 10 kcal/mol to the difference in binding energy were considered outliers and were not included in the results. These contributions are probably due to unfavourable contacts that are not resolved by the simple minimization protocol used here.

For two additional systems, a slightly different protocol gave distinctly better results. These were the BPTI:trypsin and BPTI:chymotrypsin complexes [38,39]. Several BPTI mutations at position 15 (within the interface) were considered. Instead of minimizing just the mutated sidechain, as above, we minimized the entire BPTI protein for each choice of rotamer (for 50 steps, as above).

Protein design

The energy function for protein design corresponds to Equation (1), with the solvation contribution described

by a CASA term. The interaction energy between each possible combination of sidechain pairs, or between a sidechain and the backbone, are precomputed and stored in an energy matrix. For a given sidechain pair, this calculation includes all possible combinations of both amino acid types and rotamer values. Once the energy matrix is computed, the amino acid sequence is optimized in a second stage, through cycles of random mutations and steepest-descent minimization. This heuristic procedure was developed and validated by Wernisch et al. [24]. A "heuristic cycle" proceeds as follows. An initial amino acid sequence and set of sidechain rotamers are chosen randomly. These are improved in a stepwise way. At a given amino acid position i, the best amino acid type and rotamer are selected, with the rest of the sequence and structure held fixed. The same is done for the following position $i + 1$, and so on, performing multiple passes over the amino acid sequence until the energy no longer improves (or a given, large number of passes is reached). The final sequence, rotamer set, and energy are output, ending the cycle. For the design calculations below, we performed 450,000 heuristic cycles for each protein. Disulfide-bonded cysteines, glycines and prolines are expected to have a special effect on the protein's folded and unfolded state structures, which may not be accurately captured by our method. Therefore, if these amino acids were present in the native sequence, they were held fixed; all other amino acids were allowed to mutate freely. The calculations were done using our Proteins@Home distributed computing platform. This allows us to use the computers of several thousand volunteers in over 70 countries. Proteins@Home is based on the Berkeley Open Infrastructure for Network Computing, BOINC [65]. The Proteins@Home platform and project will be described in detail elsewhere [49].

The model for the unfolded state of a protein is analogous to that described above, employing a collection of tripeptide structures. However, an additional, empirical correction was added to the unfolded state energies. For each amino acid type, the correction is defined and optimized to provide a realistic overall amino acid composition of the resulting sequences. More details of this procedure are given elsewhere [48]. The precise values of the reference energies are given in Table 8. The dielectric constant was set to $\varepsilon = 14$; the weight of the surface term (Eq. 2) was set to $\alpha = 1$. Notice that our earlier design calculations with the MF parameters [48] used $\varepsilon = 10$, $\alpha = 1$, and different reference energies.

Table 8: Reference energies (kcal/mol) characterizing the unfolded state

	initial	optimized	difference
Ala	-10.009	-11.307	1.30
Asp	-24.223	-19.826	-4.40
Asn	-20.783	-17.180	-3.60
Arg	-22.199	-25.043	2.84
Glu	-24.365	-21.257	-3.11
Gln	-20.707	-17.940	-2.77
His	-21.928	-20.389	-1.54
Ile	-13.904	-12.320	-1.58
Leu	-13.941	-12.600	-1.34
Lys	-18.946	-22.214	3.27
Met	-14.013	-13.922	-0.09
Phe	-21.741	-17.412	-4.33
Ser	-16.656	-13.450	-3.21
Tyr	-23.727	-20.274	-3.45
Thr	-16.252	-12.583	-3.67
Trp	-23.993	-20.983	-3.01
Val	-13.338	-11.481	-1.86

Authors' contributions

MSAB: wrote software, performed calculations, analyzed data, wrote paper. AL: wrote software, performed calculations, analyzed data. NA: performed calculations, analyzed data. CB: Performed calculations, analyzed data, wrote paper. TS: designed research, wrote software, wrote paper. All the authors read and approved the final manuscript.

Additional material

Additional file 1

Calculated and experimental binding free energies. Complete details on the binding free energy changes due to point mutations in the various test proteins and peptides.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-148-S1.pdf]

Acknowledgements

We thank the many volunteers who have participated in the Proteins@Home project and contributed computer cycles used in this work. See biology.polytechnique.fr/proteinsathome for a complete list of participants. We thank the BOINC development community for testing the alpha version of Proteins@Home. We thank Alexey Aleksandrov and David Mignon for discussions and the ANR High Performance Computing program for support.

References

1. Becker O, Mackerell A Jr, Roux B, Watanabe M, Eds: *Computational Biochemistry & Biophysics* Marcel Dekker, New York; 2001.
2. Guérois R, Lopez de la Paz M, Eds: *Protein Design: Methods And Applications* Humana Press; 2007.
3. Roux B, Simonson T: **Implicit solvent models.** *Biophys Chem* 1999, **78:**1-20.
4. Eisenberg D, McClachlan A: **Solvation energy in protein folding and binding.** *Nature* 1986, **319:**199-203.
5. Wesson L, Eisenberg D: **Atomic solvation parameters applied to molecular dynamics of proteins in solution.** *Prot Sci* 1992, **1(2):**227-235.

6. Fraternali F, van Gunsteren W: **An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution.** *J Mol Biol* 1996, **256:**939-948.
7. Ferrara P, Apostolakis J, Caflisch A: **Evaluation of a fast implicit solvent model for molecular dynamics simulations.** *Proteins* 2002, **46:**24-33.
8. Koehl P, Delarue M: **Polar and nonpolar atomic environments in the protein core. Implications for folding and binding.** *Proteins* 1994, **20:**264-278.
9. Juffer AH, Eisenhaber F, Hubbard SJ, Walther D: **Comparison of atomic solvation parametric sets: Application and limitations in protein folding and binding.** *Prot Sci* 1995, **4:**2499-2509.
10. Pei J, Wang Q, Zhou J, Lai L: **Estimating protein-ligand binding free energy: Atomic solvation parameters for partition coefficient and solvation free energy calculation.** *Proteins* 2004, **57(4):**661-664.
11. Feig M, Brooks CL III: **Recent Advances in the Development and Application of Implicit Solvent Models in Biomolecule Simulations.** *Curr Opin Struct Biol* 2004, **14:**217-224.
12. Simonson T: **Macromolecular electrostatics: continuum models and their growing pains.** *Curr Opin Struct Biol* 2001, **11:**243-252.
13. Honig B, Nicholls A: **Classical electrostatics in biology and chemistry.** *Science* 1995, **268:**1144-1149.
14. Schaefer M, Sommer M, Karplus M: **pH-dependence of protein stability: absolute electrostatic free energy differences between conformations.** *J Phys Chem B* 1998, **101:**1663-1683.
15. Simonson T: **Electrostatics and dynamics of proteins.** *Rep Prog Phys* 2003, **66:**737-787.
16. Archontis G, Simonson T: **A residue-pairwise Generalized Born scheme suitable for protein design calculations.** *J Phys Chem B* 2005, **109:**22667-22673.
17. Ooi T, Oobatake M, Nemethy G, Scheraga H: **Accessible surface areas as a measure of the thermodynamic hydration parameters of peptides.** *Proc Natl Acad Sci USA* 1987, **84:**3086-3090.
18. Wang W, Lim W, Jakalian A, Wang J, Luo R, Bayly C, Kollman P: **An analysis of the interactions between the Sem-5 SH3 domain and its ligands using molecular dynamics, free energy calculations, and sequence analysis.** *J Am Chem Soc* 2001, **123:**3986-3994.
19. Hou T, Qiao X, Zhang W, Xu X: **Empirical aqueous solvation models based on accessible surface areas with implicit electrostatics.** *J Phys Chem B* 2002, **106:**11295-11304.
20. Lopes A, Aleksandrov A, Bathelt C, Archontis G, Simonson T: **Computational sidechain placement and protein mutagenesis with implicit solvent models.** *Proteins* 2007, **67:**853-867.
21. Bolon D, Mayo S: **Enzyme-like proteins by computational design.** *Proc Natl Acad Sci USA* 2001, **98:**14274-14279.
22. Liang S, Grishin N: **Effective scoring function for protein sequence design.** *Proteins* 2004, **54:**271-281.
23. Hellinga H, Richards F: **Optimal sequence selection in proteins of known structure by simulated evolution.** *Proc Natl Acad Sci USA* 1994, **91:**5803-5807.
24. Wernisch L, Héry S, Wodak S: **Automatic protein design with all atom force fields by exact and heuristic optimization.** *J Mol Biol* 2000, **301:**713-736.
25. Kuhlman B, Baker D: **Native protein sequences are close to optimal for their structures.** *Proc Natl Acad Sci USA* 2000, **97:**10383-10388.
26. Koehl P, Levitt M: **Protein topology and stability define the space of allowed sequences.** *Proc Natl Acad Sci USA* 2002, **99:**1280-1285.
27. Dantas G, Kuhlman B, Callender D, Wong M, Baker D: **A Large Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins.** *J Mol Biol* 2003, **332:**449-460.
28. Saunders C, Baker D: **Recapitulation of protein family divergence using flexible backbone protein design.** *J Mol Biol* 2005, **346:**631-644.
29. Madaoui H, Becker E, Guérois R: **Sequence search methods and scoring functions for the design of protein structures.** *Methods Mol Biol* 2006, **340:**183-206.
30. Kang SG, Saven JG: **Computational protein design: structure, function and combinatorial diversity.** *Curr Opin Chem Biol* 2007, **11:**329-334.
31. Zhou H, Zhou Y: **Stability scale and atomic solvation parameters extracted from 1023 mutation experiments.** *Proteins* 2002, **49:**483-492.
32. Lomize AL, Reibarkh MY, Pogozheva ID: **Interatomic potentials and solvation parameters from protein engineering data for buried residues.** *Prot Sci* 2002, **11(8):**1984-2000.
33. Makhatadze GI, Privalov PL: **Energetics of interactions of aromatic hydrocarbons with water.** *Biophys Chem* 1994, **50:**285-291.
34. Press W, Flannery B, Teukolsky S, Vetterling W: *Numerical Recipes* Cambridge University Press, Cambridge; 1986.
35. Rick SW, Berne BJ: **Free energy of the hydrophobic interaction from molecular dynamics simulations: The effects of solute and solvent polarizability.** *J Phys Chem B* 1997, **101:**10488-10493.
36. Huang X, Margulis CJ, Berne BJ: **Do molecules as small as neopentane induce a hydrophobic response similar to that of large hydrophobic surfaces?** *J Phys Chem B* 2003, **107:**11742-11748.
37. Gambacorti-Passerini CB, Gunby RH, Piazza R, Galietta A, Rostagno R, Scapozza L: **Molecular mechanisms of resistance to imatinib in philadelphia-chromosome-positive leukaemias.** *Lancet Oncol* 2003, **4:**75-85.
38. Almlöf M, Aqvist J, Smalas AO, Bransdal BO: **Probing the effect of point mutations at protein-protein interfaces with free energy calculations.** *Biophys J* 2006, **90:**433-442.
39. Krowarsch D, Dadlez M, Buczek O, Krokoszynska I, Smalas AO, Otlewski J: **Probing the effect of point mutations at protein-protein interfaces with free energy calculations.** *J Mol Biol* 1999, **289:**175-186.
40. Guérois R, Nielsen J, Serrano L: **Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.** *J Mol Biol* 2002, **320:**369-387.
41. Pokola N, Handel T: **Energy functions for protein design: Adjustement with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity.** *J Mol Biol* 2005, **347:**203-227.
42. Bashford D, Case D: **Generalized Born models of macromolecular solvation effects.** *Ann Rev Phys Chem* 2000, **51:**129-152.
43. Hawkins G, Cramer C, Truhlar D: **Pairwise descreening of solute charges from a dielectric medium.** *Chem Phys Lett* 1995, **246:**122-129.
44. Cornell W, Cieplak P, Bayly C, Gould I, Merz K, Ferguson D, Spellmeyer D, Fox T, Caldwell J, Kollman P: **A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.** *J Am Chem Soc* 1995, **117:**5179-5197.
45. Schaefer M, Karplus M: **A comprehensive analytical treatment of continuum electrostatics.** *J Phys Chem* 1996, **100:**1578-1599.
46. Calimet N, Schaefer M, Simonson T: **Protein molecular dynamics with the Generalized Born/ACE solvent model.** *Proteins* 2001, **45:**144-158.
47. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M: **CHARMM: a program for macromolecular energy, minimization, and molecular dynamics calculations.** *J Comp Chem* 1983, **4:**187-217.
48. Schmidt am Busch M, Lopes A, Mignon D, Simonson T: **Computational protein design: software implementation, parameter optimization, and performance of a simple model.** *J Comp Chem* 2007 in press.
49. Simonson T, Mignon D, Schmidt am Busch M, Lopes A, Bathelt C: **The inverse protein folding problem: structure prediction in the genomic era.** In *Distributed & Grid Computing – Science Made Transparent for Everyone. Principles, Applications and Supporting Communities* Tektum Publishers, Berlin; 2007.
50. Jaramillo A, Wernisch L, Héry S, Wodak S: **Folding free energy function selects native-like protein sequences in the core but not on the surface.** *Proc Natl Acad Sci USA* 2002, **99:**13554-13559.
51. Larson S, Garg A, Desjarlais J, Pande V: **Increased detection of structural templates using alignments of designed sequences.** *Proteins* 2003, **51:**390-396.
52. Lee B, Richards F: **The interpretation of protein structures: estimation of static accessibility.** *J Mol Biol* 1971, **55:**379-400.
53. Brünger AT: *X-PLOR version 3.1, A System for X-ray crystallography and NMR* Yale University Press, New Haven; 1992.
54. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P: **The Protein Data Bank.** *Nucl Acids Res* 2000, **28:**235-242.

55. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-Pdb-Viewer: An environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.
56. Tuffery P, Etchebest C, Hazout S, Lavery R: **A New Approach to the Rapid Determination of Protein Side Chain Conformations.** *J Biomol Struct Dyn* 1991, **8:**1267.
57. Kumar M, Bava K, Gromiha M, Parabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.** *Nucl Acids Res* 2006, **34:**D204-206.
58. Myers JK, Pace CN, Scholtz JM: **Helix propensities are identical in proteins and peptides.** *Biochemistry* 1997, **36:**10923-10929.
59. Park SH, Shalongo W, Stellwagen E: **Residue helix parameters obtained from dichroic analysis of peptides of defined sequence.** *Biochemistry* 1993, **32:**7048-7053.
60. Yang J, Spek EJ, Gong Y, Zhou H, Kallenbach NR: **The role of context on alpha-helix stabilization: host-guest analysis in a mixed background peptide model.** *Prot Sci* 1997, **6(6):**1-9.
61. Varadarajan R, Connelly PR, Sturtevant JM, Richards FM: **Heat capacity changes for protein-peptide interactions in the ribonuclease S system.** *Biochemistry* 1992, **31:**1421-1426.
62. Padmanabhan S, Marqusee S, Ridgeway T, Laue TM, Baldwin RL: **Relative helix-forming tendencies of nonpolar amino acids.** *Nature* 1990, **344:**268-270.
63. Lyu CP, Liff MI, Marky LA, Kallenbach NR: **Side chain contributions to the stability of alpha-helical structure in peptides.** *Science* 1990, **250:**669-673.
64. Shoemaker KR, Kim PS, Brems DN, Marqusee S, York EJ, Chaiken IM, Stewart JM, Baldwin RL: **Nature of the charged-group effect on the stability of the C-peptide helix.** *Proc Natl Acad Sci USA* 1985, **82:**2349-2353.
65. Anderson DP: **BOINC: A System for Public-Resource Computing and Storage.** In *5th IEEE/ACM International Workshop on Grid Computing* IEEE Computer Society Press, USA; 2004.
66. Ho CK, Fersht AR: **Internal thermodynamics of position 51 mutants and natural variants of tyrosyl-tRNA synthetase.** *Biochemistry* 1986, **25:**1891-1897.
67. Wells TN, Fersht AR: **Use of binding energy in catalysis analyzed by mutagenesis of the tyrosyl-tRNA synthetase.** *Biochemistry* 1986, **25:**1881-1886.
68. First EA, Fersht AR: **Mutational and kinetic analysis of a mobile loop in tyrosyl-tRNA synthetase.** *Biochemistry* 1993, **32:**13658-13663.
69. De Prat Gay G, Duckworth HW, Fersht AR: **Modification of the amino acid specificity of tyrosyl-tRNA synthetase by protein engineering.** *FEBS Letters* 1993, **318:**167-171.
70. Fersht AR, Leatherbarrow RJ, Wells TN: **Structure-reactivity relationships in engineered proteins:analysis of use of binding energy by linear free energy relationships.** *Biochemistry* 1987, **26:**6030-6038.
71. Sharp KA: **Calculation of HyHel10-lysozyme binding free energy changes: Effect of ten point mutations.** *Proteins* 1998, **33:**39-48.
72. Moebius U, Clayton LK, Abraham S, Harrison SC, Reinherz EL: **The human immunodeficiency virus gp120 binding site on CD4: Delineation by quantitative equilibrium and kinetic binding studies of mutants in conjunction with a high-resolution CD4 atomic structure.** *J Exp Med* 1992, **176:**507-517.
73. Cavarelli J, Eriani G, Rees B, Ruff M, Boeglin M, Mitschler A, Martin F, Gangloff J, Thierry J, Moras D: **The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction.** *EMBO J* 1994, **13:**327-337.

Ces deux études (articles *PROTEINS* et *BMC Bioinformatics*) nous ont donc permis de mettre en place une fonction d'énergie relativement simple mais néanmoins robuste puisqu'elle a fait ses preuves sous réserve de légers ajustements, dans des applications très diverses. Plus récemment, notre procédure de CPD fut utilisée pour le *design* complet de 16 protéines globulaires [185]. Le score d'identité des séquences obtenues par notre procédure pour 15 des 16 protéines étudiées est comparable à celui d'homologues naturels. De plus, la stabilité des modèles prédits fut validée par des simulations de dynamique moléculaire, les structures modélisées se montrant très similaires aux structures expérimentales. Cet article est présenté en annexe 1. L'ensemble de ces résultats est donc très encourageant pour aborder l'évolution dirigée de l'AsnRS.

Chapitre 4

Ajustement et évaluation de la fonction d'énergie pour l'évolution dirigée des synthétases

Notre but ultime est de réaliser le *design* du site actif de l'AsnRS de manière à ce qu'elle lie préférentiellement l'aspartate au détriment de son acide aminé natif, l'asparagine. Ce travail sera mené sur l'AsnRS de *T. thermophilus* dont la structure cristallographique fut déterminée par Berthet-Colominas et coll [12]. Une étude préliminaire fut nécessaire afin de réaliser des tests et ajustements de notre fonction d'énergie. En effet, nos études précédentes nous ont montré que d'une application à l'autre et d'un système à l'autre de tels ajustements étaient nécessaires même si la procédure générale était maintenue. De plus, nos paramètres d'énergie et en particulier nos énergies de références furent optimisés à l'origine sur des protéines globulaires en faisant varier à la fois des régions du coeur hydrophobe et de surface. Le site actif de l'AsnRS s'avère plus complexe puisqu'il est régi par des contraintes structurales devant aussi répondre à une nécessité fonctionnelle. Nous avons donc affiné notre fonction d'énergie à partir du système bien connu de l'AspRS d'*E. coli* qui présente de grandes similitudes avec l'AsnRS. En particulier, ce système fut utilisé pour évaluer notre procédure de CPD dans sa capacité à reproduire des séquences

à caractère natif ou à prédire des changements d'affinité associés à des mutations ponctuelles.

4.1 Matériel et méthodes

4.1.1 Construction du système

Construction des modèles d'AspRS et AsnRS

Nous avons utilisé les structures d'AspRS d'*E. coli* et d'AsnRS de *T. thermophilus*, [154], [12]. Pour chacun des systèmes nous avons conservé l'ion Mg-1 qui était présent dans chacune des structures. Par ailleurs, dans le système de l'AspRS nous avons conservé la molécule d'eau observée dans plusieurs structures cristallographiques d'AspRS. Cette molécule d'eau interagit simultanément avec l'Asp 233 et le groupement ammonium du ligand. L'AspRS et l'AsnRS sont des dimères de deux fois 550 et 438 résidus respectivement. D'un point de vue computationnel, il était impensable de réaliser le *design* des protéines entières. Notre but étant de modifier la spécificité de l'enzyme, nous nous sommes donc concentrés sur le site actif. Pour cela, nous avons défini une sphère de 15 Å autour du ligand dans laquelle les résidus sont autorisés à muter. Par la suite nous désignerons ces résidus comme 'Actifs'. Fut définie une seconde sphère de 30 Å de rayon au delà de laquelle les atomes ne sont plus pris en considération. Entre ces deux sphères, réside une zone tampon dans laquelle les résidus ne sont pas autorisés à muter. Cependant, leurs chaînes latérales ne sont pas fixées et peuvent échantilloner tous leurs rotamères respectifs (figure 4.1). Nous désignerons ces résidus comme 'Inactifs'. Enfin, les glycines, prolines et cystéines sont 'Gelées'. Ces acides aminés peuvent avoir des effets sur le repliement de la protéine difficiles à prendre en compte par notre méthode. Leur mutation pourrait donc avoir des conséquences notables sur la structure de la protéine. Par conséquent, ces résidus sont fixés en séquence et en structure indépendamment de la sphère à laquelle ils appartiennent.

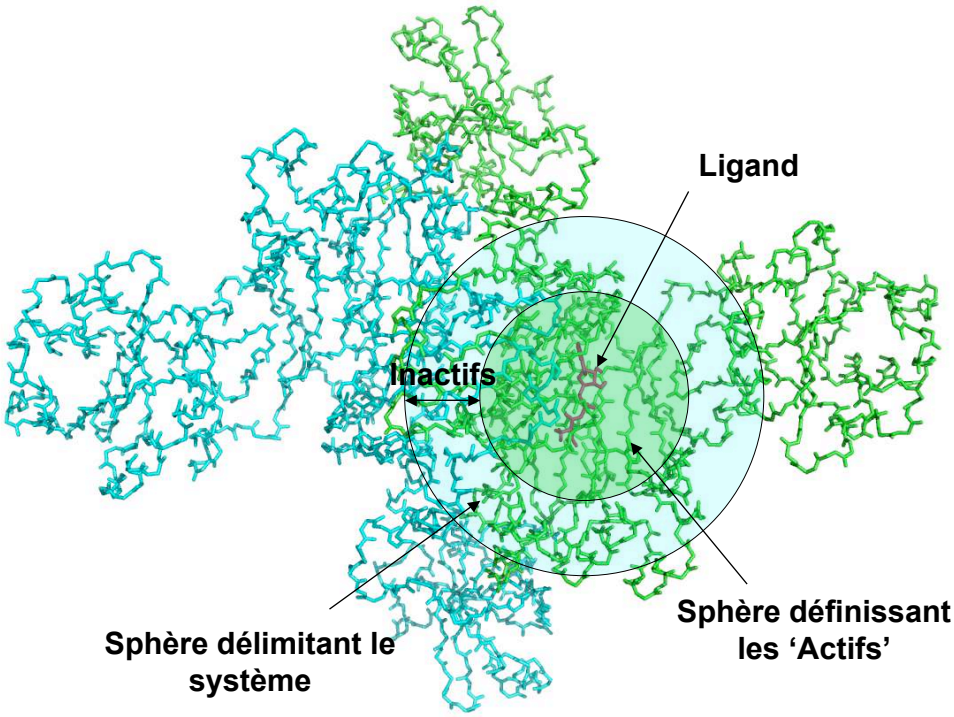


FIG. 4.1 – **Représentation du système d'étude.** Ici, est schématisé le système de l'AspRS. Le ligand est représenté en magenta. Une première sphère de 15 Å autour du ligand défini les résidus dits 'Actifs'. Une seconde sphère de 30 Å délimite le système d'étude, les résidus en dehors de la seconde sphère ne sont pas pris en compte lors du calcul de la matrice d'énergie. La zone tampon entre les deux sphères défini les résidus 'Inactifs'.

L'Arg 217 [208], l'Asp 475 [352], le Glu 482 [361] et l'Arg 489 [368] jouent un rôle déterminant dans la reconnaissance de l'ATP et du substrat (numérotation correspondant respectivement à l'AspRS d' *E. coli* et [l'AsnRS de *T. thermophilus*]). Ces positions ne sont donc pas autorisées à muter, bien que présentes dans le site actif. Toutefois, leurs chaînes latérales ne sont pas fixées, elles sont 'Inactives'. Finalement, un total de 40 et 38 résidus 'Actifs' fut défini respectivement pour le système de l'AspRS et celui de l'AsnRS.

4.1.2 Génération des rotamères du ligand

Nos calculs de CPD furent réalisés en présence d'AspAMP ou d'AsnAMP. Nous avons voulu autoriser différentes conformations pour le ligand. Il fallut donc générer des rotamères du ligand afin de mimer sa flexibilité. Pour cela, nous sommes partis de la structure cristallographique de l'AspAMP dans la poche de l'AspRS et de l'AsnAMP dans la poche de l'AsnRS. Pour chacun des deux ligands nous avons procédé comme suit. L'ensemble de la procédure est schématisé en figure 4.2.

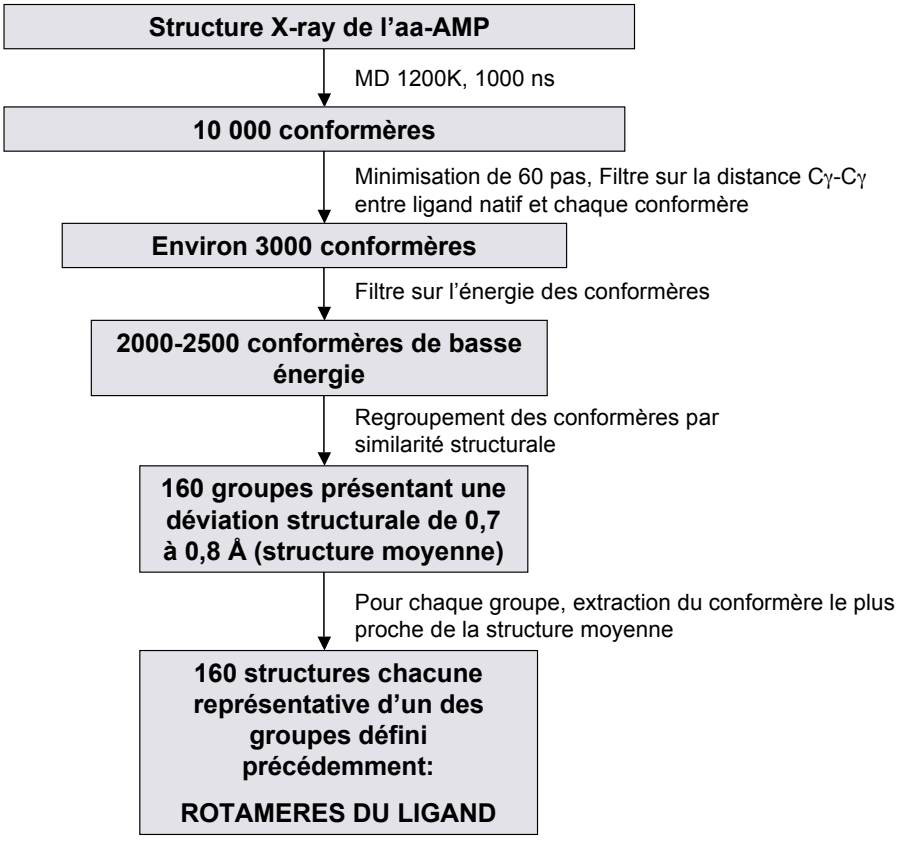


FIG. 4.2 – **Schéma de la procédure de génération des rotamères du ligand.**

-Une dynamique moléculaire est réalisée avec le programme Xplor à 1200K sur une période de 1000 ns [20]. En chauffant, on espère échantillonner un maximum de conformations du ligand. Dans le souci de préserver l'orientation générale du ligand natif, la partie commune du ligand (région AMP) est maintenue fixée (figure 4.3). On récupère les structures toutes les 100 ps pour un total de 10000 conformations.

-Provenant d'une dynamique à 1200K, les conformations ne sont pas optimales énergétiquement. Chaque structure subit donc une minimisation de 60 pas.

-Les 10000 structures sont alors filtrées sur leur orientation globale par rapport au ligand natif. Seules les structures dont le Cγ est à moins de 3 Å du Cγ natif sont retenues.

-Un second filtre est appliqué pour éliminer les structures de trop haute énergie. La structure de plus basse énergie est prise comme référence et toutes les conformations dont l'énergie lui est supérieure de plus de 10 kcal/mol sont éliminées.

-Les conformations restantes sont ensuite regroupées par similarité structurale. La déviation structurale autorisée entre chaque membre d'un groupe donné est définie par une valeur seuil. Différentes valeurs entre 0,5 Å et 1 Å furent testées. Un seuil trop fin conduit à un nombre de groupes trop important (environ 700 groupes avec 0.5 Å) tandis qu'une valeur de 1 Å donne trop peu de groupes. Des seuils de 0,7 Å pour l'AspAMP et 0,8 Å pour l'AsnAMP furent retenus (environ 160 groupes).

-Les centres des groupes correspondent à des structures moyennes avec par conséquent des géométries irréalistes. Les structures appartenant à un groupe donné sont alors comparées au centre du groupe ; on retient la structure présentant la déviation structurale la plus faible avec le centre du groupe. Ces structures sont choisies comme rotamères du ligand. Un total de 161 rotamères furent ainsi construits pour l'AspAMP et 163 pour l'AsnAMP (figure 4.3).

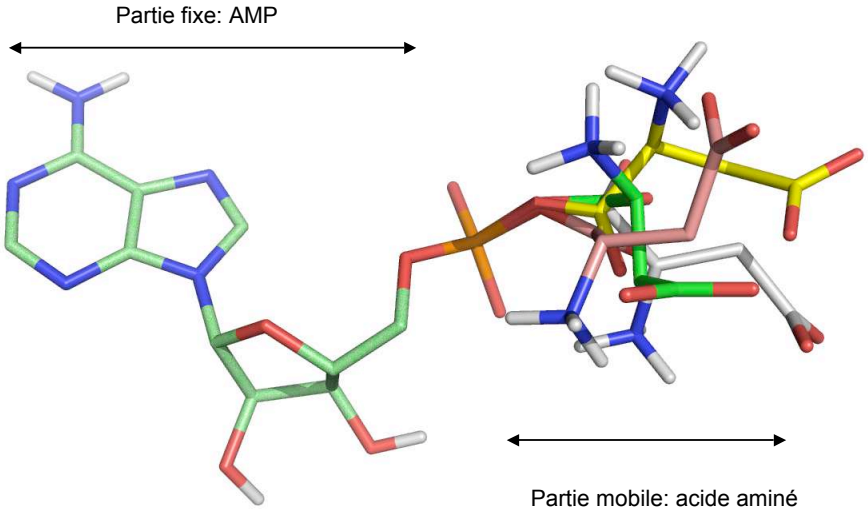


FIG. 4.3 – **Représentation de différents rotamères du ligand AspAMP.** La partie AMP est commune à tous les rotamères du ligand. Seule la région de l'acide aminé est autorisée à bouger. Le cas de l'AsnAMP est similaire (non montré).

Les couples de dièdres $\chi 1$, $\chi 2$ de chacun des rotamères de l'AspAMP et l'AsnAMP sont représentés en figure 4.4. On voit que notre ensemble de rotamères inclut bien les 5 rotamères de l'aspartate et les 11 rotamères de l'asparagine présents dans la bibliothèque de Tufféry [201]. (figure 4.4).

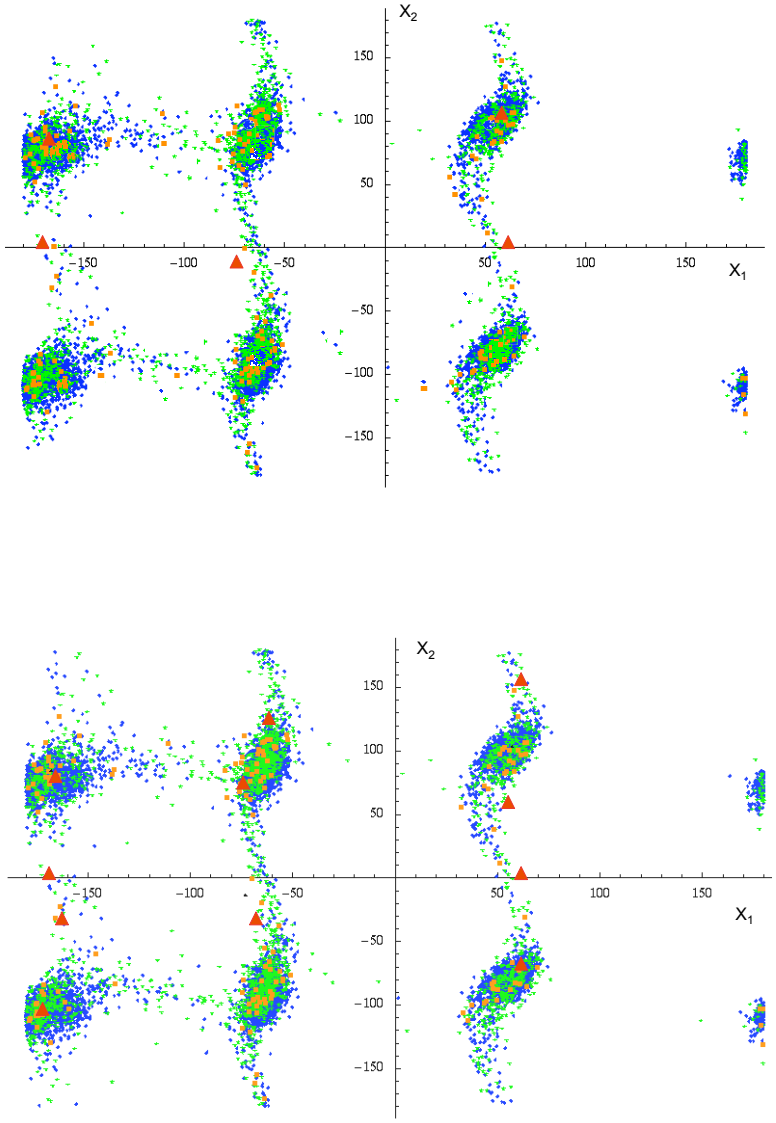


FIG. 4.4 – **Distribution des angles dièdres des conformères d'AspAMP (haut) et d'AsnAMP (bas) obtenus au cours des différentes étapes.** Bleu : les 10000 conformères de départ. Vert : les 2000-3000 conformères de basse énergie. Orange : les 161/163 rotamères finaux. Rouge : les 5/11 rotamères issus de la bibliothèque de Tufféry. Nous remarquons que tous les rotamères provenant de la librairie de Tufféry sont représentés par notre ensemble.

4.1.3 Traitement de la molécule d'eau

Poterszman et coll. rapportent la présence d'une molécule d'eau dans le site actif de l'AspRS de *T. thermophilus*. Cette molécule d'eau interagit avec l'Asp 239 [233], résidu conservé dans toutes les synthétases de classe II. Nous avons donc inclus cette molécule d'eau dans nos calculs sous la forme d'un dipôle. Des 'rotamères' du dipôle furent générés. L'extrémité négative du dipôle correspond aux coordonnées de l'oxygène de la molécule d'eau. 14 dipôles différents furent ainsi définis et son représentés en figure 4.5.

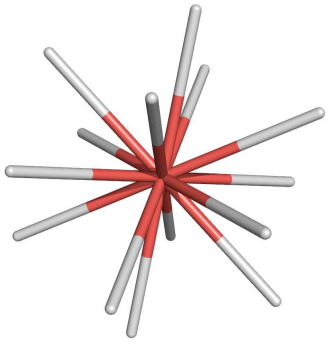


FIG. 4.5 – **Les 14 'rotamères' du dipôle.** Le centre correspond aux coordonnées cristallographiques de l'oxygène.

4.1.4 Calculs de CPD

Génération de la matrice d'énergie

La première étape du CPD consiste à calculer la matrice d'énergie qui contiendra les interactions entre toutes les paires de résidus de la protéine ('Actifs' et 'Inactifs'), en autorisant successivement tous les types d'acides aminés possibles ('Actifs') et tous les rotamères possibles ('Actifs' et 'Inactifs'). La figure 4.6 illustre le calcul d'une matrice d'énergie pour une 'protéine' de trois résidus.

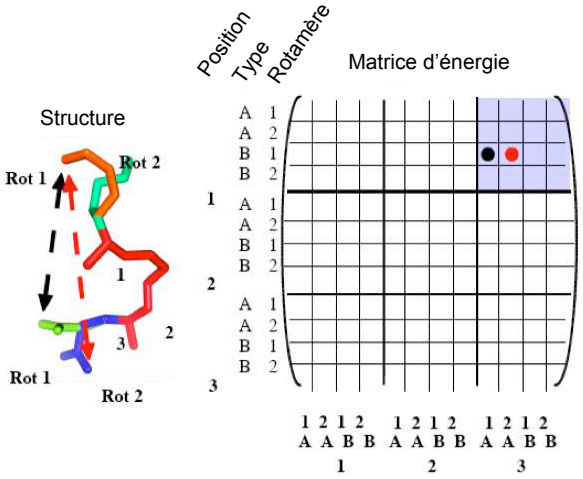


FIG. 4.6 – **Exemple de calcul de matrice d'énergie pour une 'protéine' de 3 résidus : A gauche :** La protéine en représentation 'sticks'. Le squelette peptidique est coloré en rouge. Les positions 1 et 2 possèdent chacune deux rotamères possibles, Rot1 et Rot2. On calcule successivement l'interaction entre les deux acides aminés pour toutes les combinaisons possibles de rotamères. **A droite :** les énergies résultantes sont rangées dans une matrice. Les énergies impliquant les rotamères de gauche seront stockées dans le carré gris. Les points rouge et noir correspondent respectivement aux interactions (flèche rouge, flèche noir) à gauche.

Les atomes du squelette (N, H, Cα, C et O) sont maintenus fixés pendant toutes les étapes de calculs. Les atomes des chaînes latérales sont construits géométriquement à partir des cordonnées des atomes du squelette. Les angles dièdres des chaînes latérales sont issus de la bibliothèque de rotamères de Tufféry [201]. Les calculs d'énergie sont réalisés avec le logiciel Xplor [20]. Notre fonction d'énergie décrite en sections 3.1 et 3.2 repose sur le champ de force CHARMM19 [17] et prend la forme suivante :

$$E = E_{bonds} + E_{angles} + E_{dihe} + E_{impr} + E_{vdw} + E_{coul} + E_{solv} \qquad (4.1)$$

Le solvant est décrit de façon implicite par notre modèle CASA/PHIA décrit en sections 3.1.2 et 3.2.1. On rappelle que le terme de solvatation peut être représenté par l'équation suivante :

$$E_{solv} = (\frac{1}{\epsilon} - 1)E_{coul} + \alpha \sum_i \sigma_i A_i. \tag{4.2}$$

E_{coul} représente l'énergie de Coulomb classique et ϵ correspond à la constante diélectrique du milieu. Le terme de droite correspond à la somme des surfaces accessibles au solvant A_i de tous les atomes i de la protéine, pondérées par leur paramètres de solvatation atomique respectifs (σ_i). Ce terme est lui même pondéré par le facteur α décrit en section 3.1 [134]. Pour l'ensemble de nos calculs, le facteur α a une valeur de 0.5. Les surfaces sont calculées par l'algorithme de Lee et Richard en utilisant une sphère de 1.5 Å de rayon [130]. Reprenant l'approximation de Street et Mayo, la surface accessible au solvant A_i d'un atome i est obtenue en soustrayant la surface inaccessible au solvant de ce même atome, de sa surface totale [193]. Par ailleurs, la surface inaccessible au solvant de cet atome i correspond à la somme des surfaces de contacts A_{ij} entre cet atome i et ses atomes voisins j. Cette approximation présente l'avantage que l'énergie de surface devient alors décomposable en une somme d'énergies de paires d'acide aminés. Cependant ceci peut conduire à un sur-comptage systématique des surfaces de contacts, une portion d'un atome i pouvant être en contact avec deux atomes j et j' à la fois. Reprenant la correction empirique de Street et Mayo, nous avons montré que cette erreur pouvait être corrigée en pondérant par un facteur de 0.5 les surfaces de contacts A_{ij} impliquant au moins un atome enfoui [134].

Le calcul des énergies d'interaction entre les paires de résidus ou entre un résidu et le squelette peptidique implique une légère minimisation des chaînes latérales afin d'éliminer les conflits stériques provenant de la représentation discrète de l'espace conformationnel. Cette étape étant très coûteuse, seules les interactions entre les résidus dont les Cβ sont distants de moins de 15 Å sont calculées. Chaque chaîne latérale subit une minimisation de Powell de 15 pas en présence du squelette peptidique uniquement, afin d'obtenir son énergie 'propre'. Puis, les énergies d'interaction entre paires de résidus sont obtenues après minimisation de 30 pas des paires de chaînes latérales. L'énergie intra-squelette peptidique s'annule dans la mesure où le squelette est fixé lors de nos calculs. Ces énergies sont alors stockées dans la matrice décrite ci dessus (figure 4.6).

Description de l'état déplié

Les combinaisons séquences/structures seront évaluées selon leur énergie libre de repliement $\Delta G_{repliement}$ qui correspond à la différence d'énergie libre entre l'état replié et l'état déplié. Dans l'état replié, les coordonnées du squelette peptidique sont maintenues fixes et proviennent de la structure cristallographique. L'état déplié est décrit par le modèle tripeptide décrit en sections 1.1.2 et 3.2.1. L'énergie de référence, E_X d'un acide aminé X représente la contribution de X à l'énergie libre de l'état déplié. Ces énergies furent corrigées empiriquement de façon à obtenir des compositions d'acides aminés raisonnables pour un jeu test de quatre protéines appartenant à la famille des domaines SH3 [185]. Ces énergies de références corrigées sont présentées table 4.1.

Acide aminé	E_{ref} initiale	E_{ref} finale	correction
Ala	-10.00	-11.31	-1.31
Asp	-24.22	-19.82	4.39
Asn	-20.78	-17.18	3.60
Arg	-22.20	-25.04	2.83
Glu	-24.36	-21.25	-4.99
Gln	-20.70	-17.94	3.10
His	-21.92	-20.38	1.54
Ile	-13.90	-12.32	-1.58
Leu	-13.94	-12.60	1.34
Lys	-18.94	-22.21	-3.26
Met	-14.01	-13.92	0.09
Phe	-21.74	-17.41	4.33
Ser	-16.65	-13.45	3.19
Tyr	-23.72	-20.27	3.45
Thr	-16.25	-12.58	3.67
Trp	-23.99	-20.98	3.01
Val	-13.33	-11.48	-1.85

Table 4.1 : Correction empirique des énergies de références (kcal/mol). Le protocole détaillé est décrit dans l'article de Schmidt et coll. [185]. Ces énergies ont été calculées avec le modèle de solvant CASA/PHIA.

Optimisation des séquences

Pour cette étape, nous avons utilisé la procédure heuristique de Wernisch et coll. [211]. La procédure est décrite en détail en section 1.3.1. L'avantage de cet algorithme est qu'en multipliant les points de départ aléatoires, il permet en un temps réduit, d'explorer un espace des conformations et des séquences plus important qu'avec une

procédure déterministe. Lors de nos calculs de CPD, un total de 250 000 cycles heuristiques sont réalisés. Pour chacun des cycles, la meilleure séquence est retenue, conduisant ainsi à un total de 250 000 séquences. Un premier filtre éliminant les séquences redondantes est alors appliqué. Puis les séquences peuvent être classées selon leur score énergétique, leur score BLOSUM ou encore les deux simultanément. La procédure est implémentée dans un programme C/C++ appelée Proteus. La matrice de similarité BLOSUM utilisée est donnée dans la figure 4.7.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

FIG. 4.7 – **Matrice de score BLOSUM62 utilisée pour le classement des séquences prédites par notre algorithme.**

Reconstruction des modèles 3D

Pour chaque position, Proteus identifie l'acide aminé optimal dans sa meilleure conformation. De même, il détermine le meilleur rotamère du ligand. Cette information est stockée dans un fichier de sortie qui contient pour chaque cycle heuristique, la combinaison séquence/rotamères de plus basse énergie (figure 4.8). Par soucis de concision, nous parlerons de 'séquences de basse énergie'. A partir d'une telle séquence, un modèle 3D peut être reconstruit. En effet, les coordonnées du sque-

lette peptidique sont maintenues fixées durant toute la procédure. Il suffit donc de positionner chaque acide aminé dans le rotamère prédit par Proteus. Les atomes de la chaîne latérale sont construits géométriquement à partir des coordonnées des atomes du squelette et de celles des rotamères provenant de la librairie de Tufféry. On réalise ensuite une minimisation de 200 pas afin de réajuster l'ensemble des chaînes latérales et ainsi éliminer les conflits stériques découlant de la discrétisation de l'espace conformationnel. Le squelette peptidique et la région AMP du ligand demeurent fixés pendant cette minimisation.

```
> seq 1 AspRS score BLOSUM 94     --> varied informations (seq number, BLOSUM score...)
AA/ A M L F K F T                 --> amino acid sequence
SEQ/ 1 3 4 7 8 9 11               --> position
ROT/ 1 9 4 7 9 8 1                --> best rotamer at the different positions
```

FIG. 4.8 – **Exemple de fichier de sortie Proteus.** L'utilisateur a le choix de différents formats de sortie. L'étape de reconstruction requiert un format contenant la séquence d'acides aminés optimale (au format fasta) mais aussi les positions (numérotation PDB) et rotamères correspondants.

4.1.5 Prédiction de changements d'affinité associés à des mutations ponctuelles

Afin d'évaluer notre fonction d'énergie et notre procédure de CPD, nous nous sommes attelés à la prédiction de changement d'affinité protéine-ligand associés à des mutations ponctuelles. En particulier nous avons réalisé des mutations dans le site actif de l'AspRS pour lesquelles des données expérimentales étaient disponibles [26]. Deux protocoles furent utilisés pour la prédiction des énergies de liaison des complexes natifs et mutants.

Protocole 'minimal'

Ce protocole est aussi décrit dans l'article de Schmidt et coll. section 3.2. L'énergie de liaison protéine-ligand est décrite par l'équation suivante :

$$E_{liaison} = E_{complx} - (E_{prot} + E_{lig}). \tag{4.3}$$

E_{complx} correspond à l'énergie du complexe protéine-ligand tandis que E_{prot} et E_{lig} représentent respectivement l'énergie de la protéine et du ligand dans leur forme libre. Dans le complexe, on considère successivement tous les rotamères de la chaîne latérale mutée. Pour chaque rotamère, une minimisation de 50 pas est effectuée en présence du ligand, le reste de la protéine étant maintenu fixé. Dans la forme libre, le ligand est éloigné de la protéine et les rotamères de l'acide aminé muté sont évalués de la même façon. Les meilleures énergies pour les systèmes complexés et libres sont alors retenues.

Pour la structure native, nous ne testons pas les différents rotamères du résidu que nous voulons muter. Nous considérons dans ce cas que son orientation est optimale. La chaîne latérale est toutefois soumise à une minimisation de 50 pas. On compare ensuite l'énergie de liaison du complexe natif et mutant.

Protocole avec optimisation globale des rotamères

Par convention, nous nommerons ce protocole, le protocole 'Proteus'. Dans cette approche, nous reconstruisons tous les rotamères du système après avoir introduit la mutation désirée. Pour ce faire, nous repartons de la matrice d'énergie calculée précédemment et nous fixons dans Proteus la séquence désirée. Ainsi, nous réduisons l'application de Proteus à une simple recherche conformationnelle, le but étant de retrouver la combinaison de rotamères optimale. Dans cette approche, tous les rotamères sont autorisés à varier, contrairement au protocole minimal où seule la chaîne latérale du résidu muté varie. Un total de 20000 cycles heuristiques est réalisé. Les 1000 combinaisons de rotamères de plus basse énergie sont alors retenues. Les structures 3D correspondantes sont reconstruites par le protocole décrit en section 4.1.4. Chacune des structures subit une minimisation de 200 pas en présence et en

absence du ligand. Le squelette peptidique est maintenu fixé pendant l'étape de minimisation. L'énergie de liaison protéine-ligand est calculée pour chacune des 1000 structures et moyennée sur l'ensemble. L'ensemble de ces étapes est répété à l'identique pour le système natif où cette fois on fixe la séquence native et on optimise les rotamères pour cette séquence. Les énergies de liaisons moyennes des complexes natifs et mutants peuvent alors être comparées.

Les algorithmes stochastiques ne garantissent pas la détermination du minimum global. De plus, du fait de la discrétisation de l'espace conformationnel et des approximations faîtes dans notre fonction d'énergie, le paysage énergétique décrit par notre champ de force ne représente pas le 'vrai' paysage énergétique. Ainsi, le complexe de plus basse énergie ne correspond pas exactement au complexe 'biologique'. Il semble donc raisonnable de retenir un ensemble de solutions d'énergie similaire pour caractériser les complexes natifs et mutants. L'expérience montre que l'ensemble de ces structures présente des énergies relativement proches (voir section 4.2.1). L'énergie de liaison du complexe est donc obtenue en moyennant les énergies de liaisons des solutions retenues.

4.2 Résultats

Nous avons continué d'affiner notre fonction d'énergie sur le système de l'AspRS en présence des ligands AspAMP et AsnAMP. En particulier nous avons testé sa capacité à prédire des séquences de type natif. Plusieurs variantes du modèle furent explorées avec différentes valeurs de la constante diélectrique ϵ, du poids α et des énergies de référence. Ensuite, nous avons testé sa capacité à reproduire des changements d'affinité associés à des mutations ponctuelles de l'AspRS.

4.2.1 Analyse des séquences obtenues

On utilise le modèle sphérique décrit en section 4.1 avec 40 résidus 'Actifs'. Les complexes AspAMP et AsnAMP seront considérés séparément. Pour chaque jeu de paramètres, un total de 250 000 cycles heuristiques est réalisé conduisant à 250 000 jeux de séquences/rotamères. Après avoir éliminé les jeux redondants, nous classons les jeux restant selon leur score énergétique ou leur score BLOSUM.

Analyse des solutions de plus basse énergie

Nous retenons les 100 solutions de plus basse énergie pour une analyse détaillée. Ces séquences ont été obtenues avec le modèle CASA/PHIA, une constante diélectrique de 16 et un facteur de pondération du terme de surface de 0.5. Les calculs d'énergie ont été effectués en présence du ligand AspAMP. Les solutions obtenues sont très proches d'un point de vue énergétique. Elles s'échantillonnent entre -2440 kcal/mol pour les meilleures et -2434 kcal/mol pour les moins bonnes. Les valeurs de ces énergies sont surprenantes, néanmoins cette observation s'explique par le fait que ces énergies ne représentent pas l'énergie réelle de repliement des protéines étudiées (le même phénomène est obtenu pour la TyrRS et l'AsnRS). En effet, dans ces énergies, la contribution intra-squelette n'est pas prise en considération. De plus nous utilisons un modèle de l'état déplié très simple qui ne prend pas en compte, même si elles sont peu nombreuses, les interactions inter-résidus. Les contributions entropiques des chaînes latérales et du squelette ne sont pas calculées non plus. Enfin nous travaillons sur un système tronqué (la sphère d'AspRS ou d'AsnRS), son énergie de repliement ne correspond donc pas à l'énergie de repliement réelle. Cependant, nous avons montré que nous étions capables de comparer et d'évaluer ces séquences les unes par rapport aux autres [186]. Par ailleurs, nous avons montré que nous étions capables de prédire l'effet de mutations sur la stabilité ou l'affinité [134], [186]. Les 30 meilleures solutions sont présentées dans la figure 4.9. D'autre part, l'alignement de positions des séquences naturelles d'AspRS et AsnRS issues de différents organismes procaryotes et eucaryotes est présenté en figure 4.10.

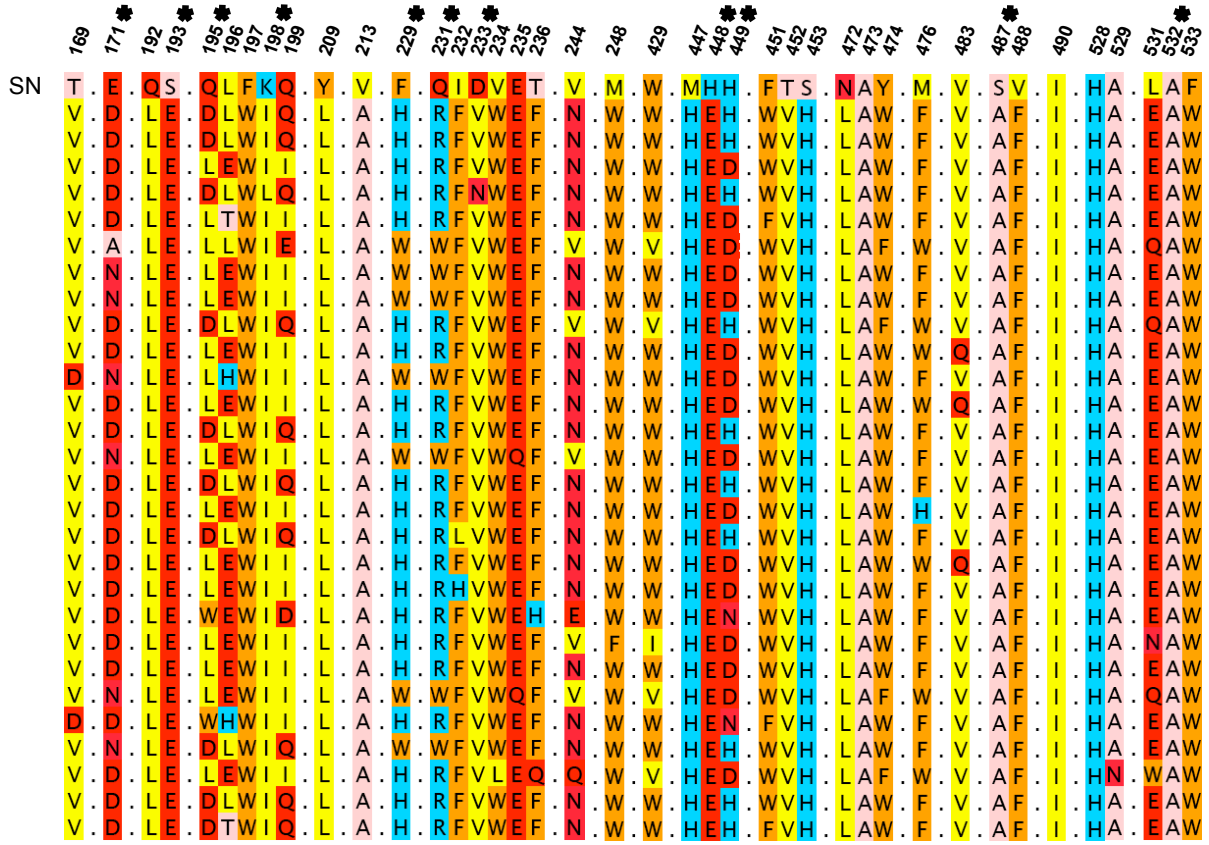


FIG. 4.9 – **Alignement de positions des séquences de meilleures énergie obtenues en présence de l'AspAMP.** Pour plus de clarté, seules les 40 positions 'Actives' sont représentées, les autres positions demeurant invariantes sont par conséquent représentées par un point '.'. Rappelons que ces 40 positions ne sont pas contiguës dans la séquence, aussi la première ligne indique leur numérotation selon la séquence d'AspRS d'*E. coli*. Les résidus à proximité du ligand sont marqués d'une étoile. La première ligne 'SN' représente la séquence native d'AspRS d'*E. coli*. Le code couleur suit la règle suivante : jaune : acides aminés hydrophobes (Val, Leu, Met et Ile), orange : acides aminés aromatiques (Phe, Trp, Tyr), rose : acides aminés de petite taille (Ala, Thr, Ser), turquoise : acides aminés chargés positivement (Arg, Lys et His protonée), rouge : acides aminés polaires et chargés négativement (Asp, Glu, Asn et Gln). L'édition des alignements a été réalisée avec l'aide de Jalview [35].

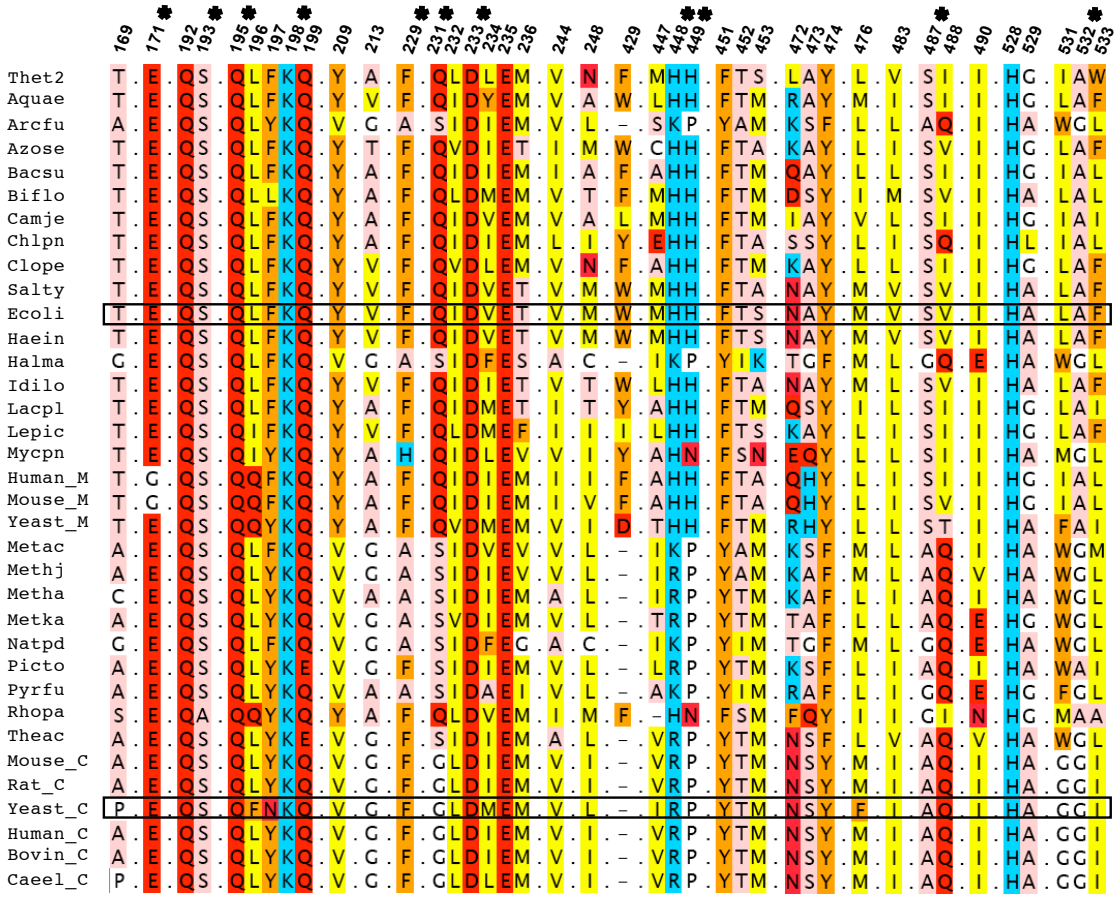


FIG. 4.10 – **Alignement de positions des séquences naturelles d'AspRS provenant de différents organismes.** Le code couleur utilisé pour la figure 4.9 reste inchangé, seules les 40 positions 'Actives' sont représentées. Les séquences d'AspRS issues d'*E. coli* et de *S. cerevisae* sont encadrées en noir.

Les séquences obtenues atteignent des scores BLOSUM de 46 par rapport à la séquence native tandis que la séquence native contre elle même conduit à un score de 210. Le score BLOSUM est calculé uniquement sur l'ensemble des positions actives. Les scores obtenus sont du même ordre de grandeur que pour certaines AspRS issues d'archaebactéries telles que Halobacteriaceae Haloarcula (SYD_HALMA (code swissprot)), Halobacteriaceae Natronomas (SYD_NATPD), Methanopyraceae Methanopyrus (SYD_METKA) ou encore Thermococcaceae Pyrococcus (SYD_PYRFU) (table 4.2). L'identité de séquence entre la séquence de plus basse énergie et la séquence native est de 25%.

	séquences	SYD_Ecoli
	SYD_Ecoli	210
	SYD_Salty	207
procaryotes	SYD_Azose	180
bactéries	SYD_Clope	165
	SYD_Thet2	164
	SYD_Chlpn	151
	SYDC_Human	92
	SYDC_Mouse	92
eucaryotes	SYDC_Ponpy	92
	SYDC_Bovin	92
	SYDC_Yeast	75
	SYD_Metka	69
procaryotes	SYD_Pyrfu	66
archae-bactéries	SYD_Halma	69
	SYD_Natpd	66

Table : 4.2 : Score BLOSUM de séquences d'AspRS provenant de différents organismes calculé contre la séquence d'AspRS d'*E. coli*. La seconde colonne indique le nom d'entrée de chacune des séquences tel qu'il apparaît dans la base de donnée 'swissprot' (http ://expasy.org). Le score BLOSUM est contenu dans la dernière colonne. Rappelons que ce dernier est calculé uniquement sur les 40 positions variables. Dans le cas des eucaryotes, 'SYDC' réfère aux AspRS cytoplasmiques. Sans surprise, les séquences mitochondriales présentent des scores BLOSUM plus élevées, les mitonchondries étant les organites très anciens possédant leur propre ADN.

L'analyse des séquences obtenues révèle une bonne prédiction des résidus de petite taille. En particulier, les trois alanines natives sont parfaitement retrouvées. La Ser 487 est systématiquement remplacée par une alanine ce qui est cohérent avec l'alignement des séquences naturelles montrant une alternance d'alanines et de serines à cette position (figure 4.10).

D'autre part, on constate que les résidus aromatiques sont prédits correctement exceptées la Tyr 209 et la Phe 229. Cette phénylalanine est généralement conservée chez les aaRS de classe II. Elle stabilise l'adénosine du ligand par le biais d'interactions van der Waals. Lorsqu'elle n'est pas conservée, elle est systématiquement remplacée par une histidine ou un tryptophane préservant ainsi un caractère aromatique. Les mutations F229H ou F229W prédites par notre algorithme, semblent maintenir ce rôle stabilisateur (figure 4.11). Par ailleurs, le remplacement systématique de la Tyr 209 par une leucine n'est pas surprenant. En effet, dans les séquences naturelles, la position 209 est souvent remplacée par une valine.

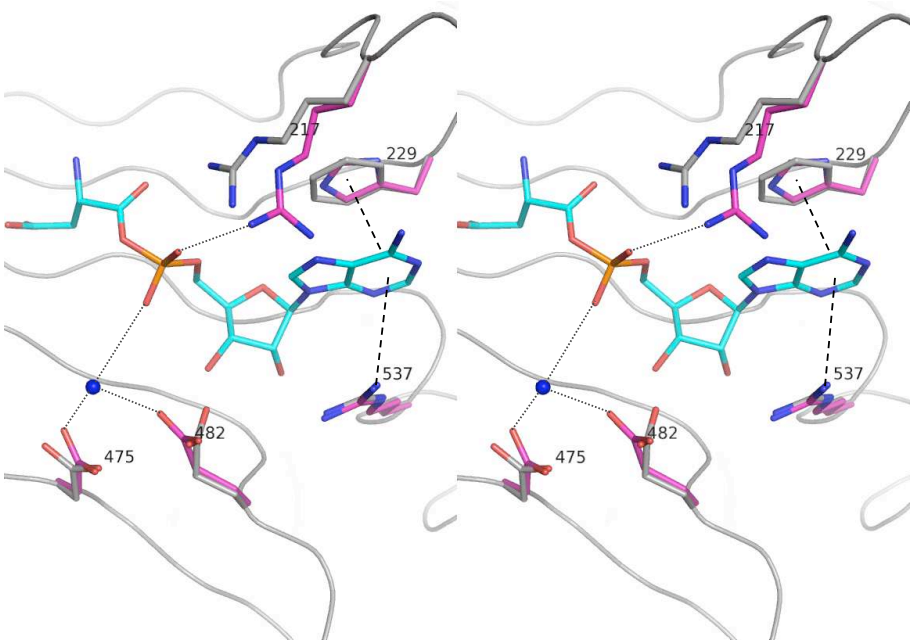


FIG. 4.11 – **Vue stéréo de la stabilisation de l'adénosine du ligand (turquoise) par la Phe 229 native (gris) ou la mutation F229H (magenta) via des interactions van der Waals.** On notera que l'Arg 537 intervenant elle aussi dans la stabilisation de l'adénosine présente la même conformation dans les structures native et mutante. De même, pour les résidus 475 et 482, les rotamères prédits sont similaires aux rotamères natifs. Le cation Mg-1 qui interagit avec le phosphate α est représenté par une sphère bleue. Les liaisons hydrogènes et ponts salins sont matérialisés par des lignes pointillées. Les interactions van der Waals sont représentées par des lignes plus épaisses. La numérotation correspond à celle de l'AspRS d'*E. coli*.

Les résidus hydrophobes non aromatiques sont eux aussi relativement bien conservés ; voir par exemple les cas de Leu 196, Val 483 et Ile 490. L'alanine remplaçant systématiquement la Val 213 dans nos séquences est cohérente avec la diversité des séquences naturelles à cette position ; la position 213 y est principalement occupée par des valines, alanines ou glycines. Nous constatons à plusieurs reprises le remplacement de résidus hydrophobes non aromatiques par des résidus hydrophobes aromatiques (V234W, I232F, M248W, M476F, V488F). Ceci peut provenir du fait que notre fonction d'énergie tend à favoriser l'enfouissement des résidus hydrophobes et en particulier des aromatiques. Ces derniers stabilisent la protéine grâce à leurs interactions van der Waals très favorables. Ainsi dès que la place est suffisante notre algorithme a tendance à introduire un résidu hydrophobe aromatique. La figure 4.12 illustre un exemple où une méthionine et une phénylalanine sont remplacées par deux tryptophanes stabilisés mutuellement par des interactions van der Waals. Ce phénomène est aussi observé dans les séquences naturelles. La Val 234 se trouve parfois mutée en tyrosine ou phénylalanine, résultat cohérent avec le tryptophane que nous prédisons à cette position.

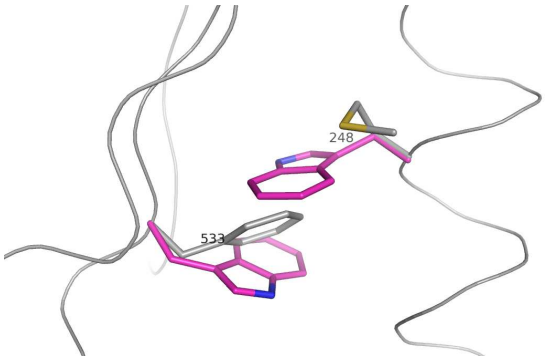


FIG. 4.12 – **Stabilisation du coeur protéique par les mutations M248W et F533W chez l'AspRS d'*E. coli*.** En gris les résidus natifs, en magenta les mutants. Les deux tryptophanes renforcent la stabilité de la protéine via des interactions van der Waals.

La prédiction des résidus polaires ou chargés s'avère plus compliquée. Les résidus polaires ou chargés enfouis ont généralement un rôle fonctionnel, phénomène difficile à prendre en compte par nos fonctions d'énergie. Il est d'ailleurs intéressant de remarquer que les régions les moins bien prédites correspondent aux résidus directement en contact avec le ligand. En particulier, on notera trois mutations remarquables : S193E, K198I et D233V. D'après les structures cristallographiques d'AspRS, la Ser 193 interagit avec le groupement ammonium du ligand [154] [26], [169]. Or, nos prédictions la remplacent systématiquement par un glutamate. L'analyse des structures prédites montre que le glutamate introduit par la mutation S193E vient aussi stabiliser le groupement ammonium du ligand dont l'orientation a été légèrement modifiée. La nouvelle orientation du ligand laisse de l'espace pour la chaîne latérale du glutamate bien plus encombrante que celle d'une serine (figure 4.13).

Par ailleurs, les résidus 198 et 233 strictement conservés chez toutes les AspRS se voient automatiquement remplacés par des résidus hydrophobes. Ces résidus jouent un rôle important dans la fonction de l'enzyme. La Lys 198 participe en effet au recrutement du ligand en interagissant avec l'od2 de l'aspartate et joue un rôle dans la discrimination Asp/Asn (voir section 2.5.1). Ce phénomène n'est pas évident à modéliser. Pour cela, il aurait fallu introduire dans notre algorithme d'optimisation, une compétition entre le ligand natif et d'autre(s) ligand(s) non natif(s) afin de rendre compte du phénomène de discrimination en faveur de l'aspartate. Cette idée est actuellement en cours d'implémentation dans notre programme Proteus. L'Asp 233 joue lui aussi un rôle fonctionnel en stabilisant le groupement ammonium de l'aspartate par l'intermédiaire d'une molécule d'eau visible dans la plupart des structures cristallographiques. L'orientation du ligand ayant été légèrement modifiée, le groupement ammonium se voit alors déplacé vers le centre du site actif et ne peut plus interagir directement avec la molécule d'eau, ni avec l'Asp 233 (figure 4.13). La

présence de l'Asp 233 n'est donc plus nécessaire. Ainsi notre fonction d'énergie va favoriser la présence d'un résidu hydrophobe à cette position enfouie pour augmenter la stabilité de la protéine.

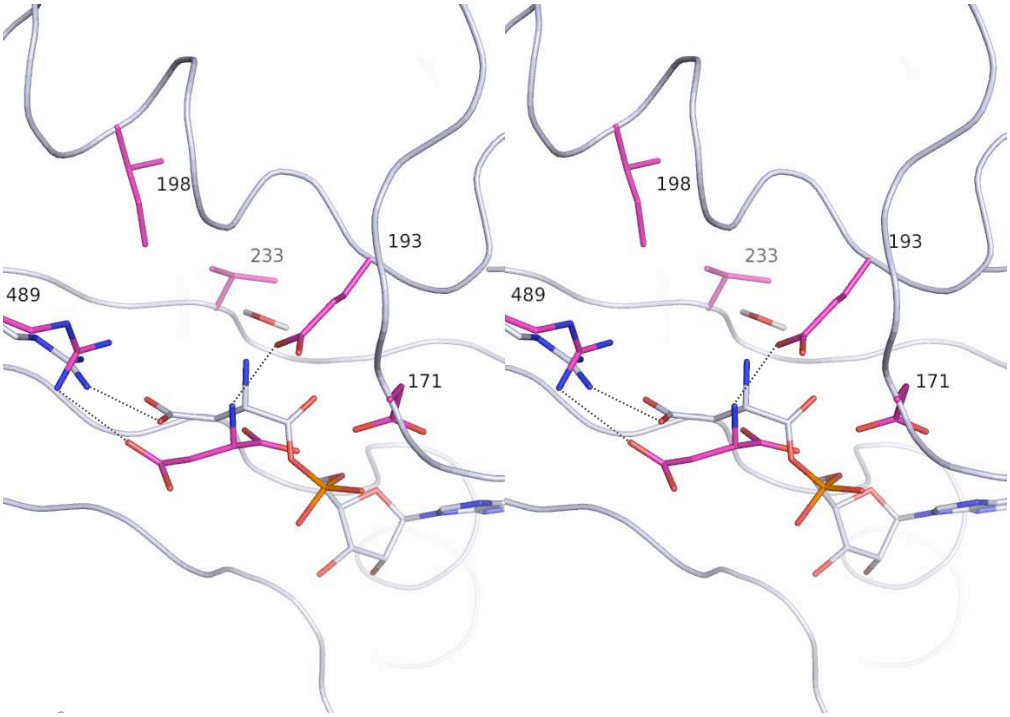


FIG. 4.13 – **Représentation stéréo du site de l'AspRS d'*E. coli* (gris) et d'un mutant pour lequel l'Asp 233 est muté en valine (magenta).** La molécule d'eau est représentée en 'sticks'. L'orientation du ligand est légèrement différente chez le mutant. Il ne peut donc plus interagir avec la molécule d'eau. Dans la structure native, le Glu 171 stabilise le groupement NH_3^+ du ligand. Chez le mutant, ce pont salin est perdu. Cette position est occupée par un aspartate qui n'interagit pas avec le groupement ammonium. Il est intéressant de voir que la mutation S193E restaure le rôle du Glu 171 en réalisant un pont salin avec le groupe NH_3^+ du ligand. On notera la performance de notre algorithme qui prédit, pour l'Arg 489, un rotamère très proche de celui observé dans la structure native. Les liaisons hydrogènes et les ponts salins sont matérialisés par des lignes pointillées.

L'Asp 233 interagit également avec la Lys 198 par l'intermédiaire d'un pont salin (figure 4.13). Sa présence dans les séquences calculées est couplée à celle de la Lys 198. Quand la Lys 198 est absente de nos prédictions, l'Asp 233 n'y apparait pas non plus. Il est intéressant de remarquer qu'à l'inverse le Glu 235, résidu strictement conservé chez l'AspRS et l'AsnRS, est parfaitement conservé dans nos prédictions. Ce dernier interagit avec l'Arg 489 que nous avons volontairement maintenu fixé en séquence. Par conséquent, nous avons indirectement orienté l'algorithme d'optimisation vers le résidu natif Glu 235.

L'ensemble de ces résultats est relativement satisfaisant. En effet, on relève un pourcentage d'identité stricte entre la séquence native et la séquence de plus basse énergie de 25% et un pourcentage d'homologie de 42,5 %. Ce résultat est proche de celui obtenu pour les domaines SH3 où l'on atteint 22,8% d'identité avec la séquence native pour les séquences de plus basse énergie (pourcentage obtenu lorsqu'on ne prend pas en considération les résidus Pro, Cys et Gly maintenus fixés durant le calcul de la matrice d'énergie) [185]. Ce résultat est d'autant plus encourageant que les énergies de références utilisées pour le système de l'AspRS furent optimisées pour les domaines SH3 et non pour les synthétases.

D'autre part, il faut souligner que nous avons ici appliqué notre procédure au *design* d'un site actif, région très riche en résidus chargés et beaucoup plus difficile à prédire. En effet, les résidus chargés du site actif et en particulier ceux directement en contact avec le ligand sont principalement régis par la nécessité de conférer une fonction précise à la protéine (reconnaissance du substrat, spécificité de l'interaction, catalyse). Or, ce phénomène est difficile à modéliser. De plus, nous utilisons un modèle de solvant très simple qui tend à favoriser l'enfouissement des résidus hydrophobes et l'exposition des hydrophiles. Il n'est donc pas surprenant que les résidus chargés enfouis dans le site actif soient moins bien prédits que les autres.

Analyse des jeux de séquences de meilleur score BLOSUM

Dans cette approche, nous classons les séquences obtenues par leur score BLOSUM62 en utilisant la séquence native comme séquence cible. Dans un premier temps, nous avons ajusté la valeur de notre constante diélectrique. Le système étudié, en particulier le site actif, est plus chargé que la plupart des systèmes que nous avons étudié au préalable. Par conséquent, la constante diélectrique optimale pour les systèmes précédents ne l'est pas nécessairement pour les sites actifs de l'AspRS ou l'AsnRS.

Nous avons donc testé différentes valeurs de constante diélectrique allant de 8 à 20. Avec des valeurs proches de 8, nous produisons des séquences beaucoup trop riches en résidus chargés. La plupart des résidus hydrophobes se voient remplacés systématiquement par des résidus polaires. Les meilleures séquences ont un score BLOSUM62 de 59. Nous rappelons que la séquence native contre elle même donne un score de 210. A l'inverse, l'utilisation d'une constante diélectrique proche de 20 donne des séquences trop hydrophobes. Le score BLOSUM des meilleures séquences est de 60. Une sélection des séquences obtenues avec des constantes diélectriques de 8 et de 16 est présentée en figure 4.14.

L'analyse des résultats obtenus montre qu'une constante diélectrique de 12 fournit un bon compromis. Les meilleures séquences présentent un score BLOSUM de 70 et une identité de séquence avec la native de 30%. Cette fois, on prédit correctement des positions hydrophobes que l'on perdait avec une constante diélectrique de 8. En effet, les positions 196, 213 et 244 retrouvent leur caractère hydrophobe (figure 4.15).

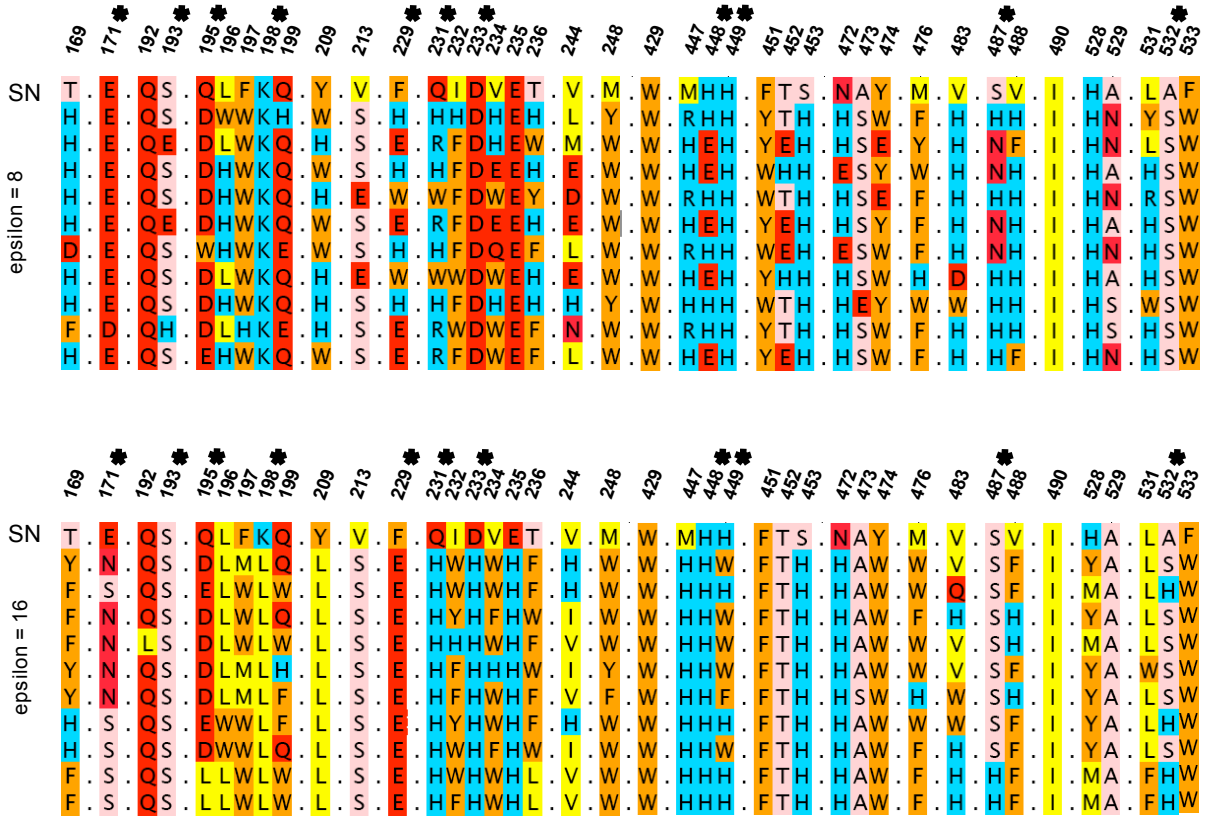


FIG. 4.14 – **Alignement de positions des séquences de meilleurs score BLOSUM obtenues en présence de l'AspAMP avec différentes valeurs de constantes diélectriques.** En haut : $\epsilon = 8$. Nous retrouvons systématiquement la lysine en position 198 et l'aspartate en position 233. Néanmoins, de nombreux hydrophobes se trouvent remplacés par des résidus chargés. En bas, $\epsilon = 16$. Ici, la position 198 est occupée par une leucine. Glu 171, essentiel au bon positionnement du ligand est remplacé par une serine. Enfin, Asp 233 et Glu 235 se trouvent convertis en histidines. On relève un excès d'hydrophobes et d'aromatiques. Comme précédemment, 'SN' indique la séquence native d'AspRS. Suivent dans chacun des cas les 10 séquences de meilleurs score BLOSUM. Le code couleur demeure inchangé (conférer figure 4.9).

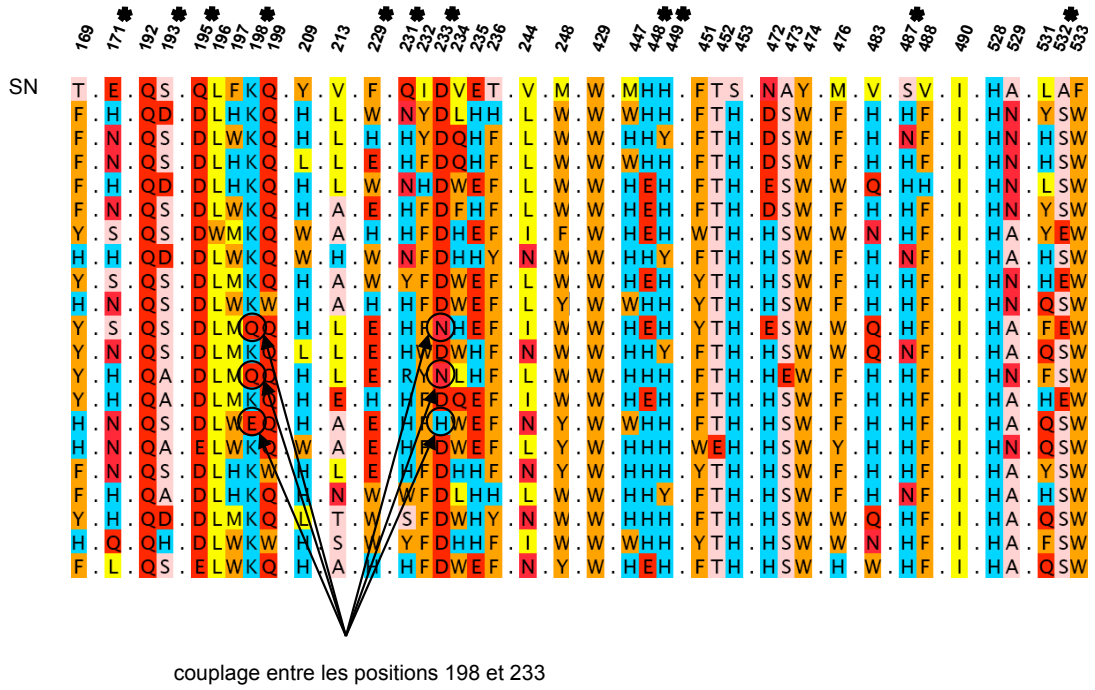


FIG. 4.15 – **Alignement de positions des séquences de meilleurs score BLOSUM obtenues en présence de l'AspAMP avec $\epsilon = 12$.** ϵ 12 semble un bon compromis entre ϵ 8 et ϵ 16. Comme précédemment, 'SN' indique la séquence native d'AspRS. Le code couleur demeure inchangé. Il est intéressant de remarquer des cas de couplage entre les positions 198 et 233. En effet la Lys 198 et l'Asp 233 apparaissent toujours ensembles. Ce couplage, aussi observé avec le protocole défini au paragraphe suivant, sera mieux détaillé par la suite.

De même, en utilisant une valeur de 12, on rétablit la lysine en position 198 ainsi que l'Asp 233 qui étaient hydrophobes lorsque ϵ=16. Néanmoins, notre fonction d'énergie tend à trop fortement favoriser les histidines et les tryptophanes. En effet, la part des tryptophanes dans nos séquences prédites atteint 15%, tandis

que les histidines occupent 25% des positions. Ce phénomène résulte probablement de l'utilisation d'énergies de références optimisées pour les domaines SH3. C'est pourquoi nous avons ajusté manuellement ces énergies de références de manière à rétablir au mieux les fréquences d'acides aminés observées chez l'AspRS et AsnRS (table 4.3).

Acide aminé	E_{ref} optm	E_{ref} corr.	Acide aminé	E_{ref} optm	E_{ref} corr.
Ala	-11.31	-10.49	Leu	-12.60	-12.26
Asp	-19.82	-19.59	Lys	-22.21	-21.33
Asn	-17.18	-16.10	Met	-13.92	-13.30
Arg	-25.04	-24.95	Phe	-17.41	-16.51
Glu	-21.25	-20.86	Ser	-13.45	-12.70
Gln	-17.94	-16.70	Tyr	-20.27	-19.06
His	-20.38	-17.98	Thr	-12.58	-12.13
Hsd	——	-21.61	Trp	-20.98	-21.00
Ile	-12.32	-12.61	Val	-11.48	-10.50

Table 4.3 : Correction empirique des énergies de références (kcal/mol) de Schmidt am Busch et coll [186]. Les valeurs sont données en kcal/mol. E_{ref} optm correspond aux énergies de référence optimisées précédemment [186] tandis que E_{ref} corr correspond aux nouvelles énergies ajustées manuellement. Nous ajoutons une nouvelle énergie de référence pour l'histidine selon qu'elle soit dans un état neutre (His) ou protoné (Hsd).

En utilisant les nouvelles énergies de références avec une constante diélectrique de 12, les meilleures séquences atteignent des scores BLOSUM supérieurs à 90 et des identités strictes de séquences de 40% (figure 4.16). Ces scores sont du même ordre que les AspRS issues d'organismes eucaryotes (table 4.2.). Il est d'ailleurs intéressant de noter que nos meilleures séquences en terme de score BLOSUM sont plus proches de la séquence native que la séquence d'AspRS cytoplasmique de levure.

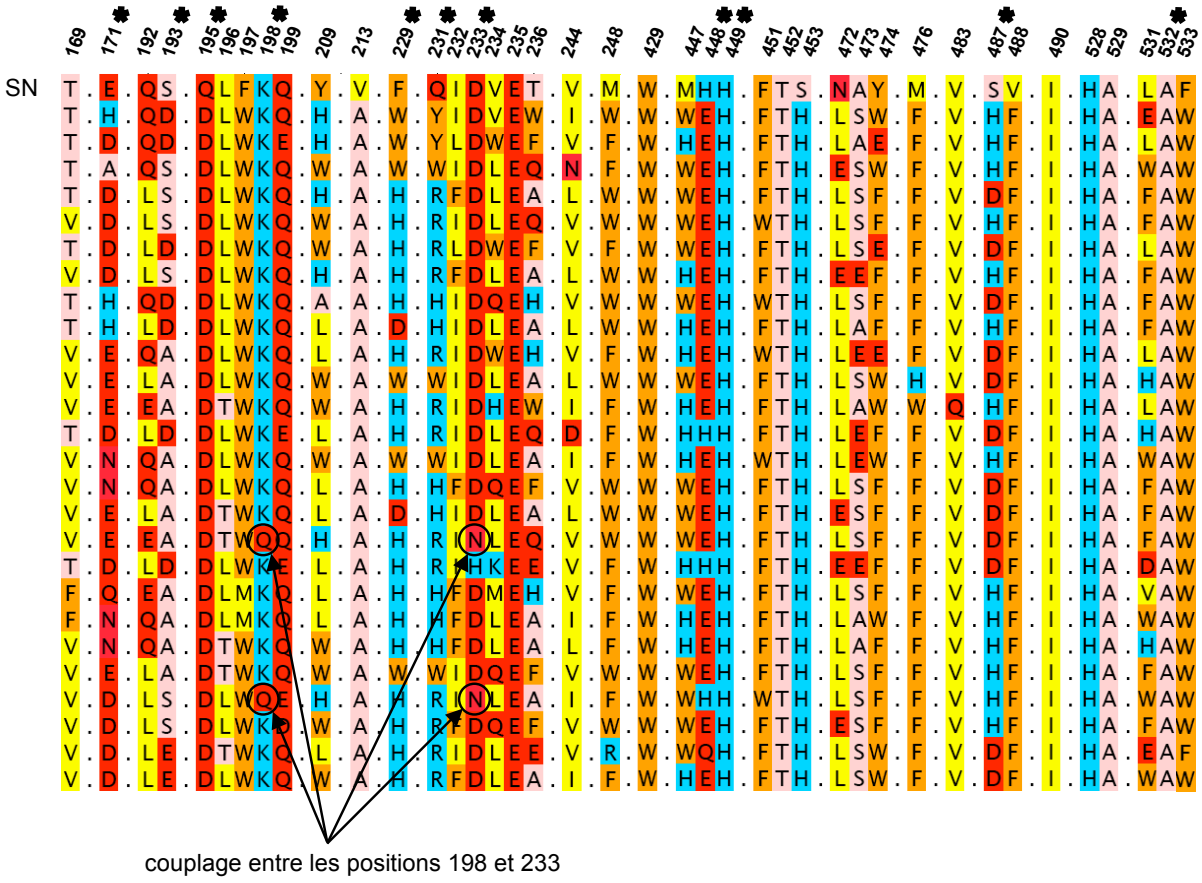


FIG. 4.16 – **Alignement de positions des séquences de meilleurs score BLOSUM obtenues en présence de l'AspAMP après optimisation des énergies de référence.** La première ligne indique la numérotation des positions selon l'AspRS d'*E. coli* tandis que 'SN' correspond à la séquence native. Les résidus à proximité du ligand sont marqués par une étoile. Le code couleur demeure inchangé. Deux cas de couplage entre les positions 198 et 233 sont observés et indiqués par des flèches noires.

Résidus conservés chez l'AspRS : Avec ce nouveau protocole, nous prédisons correctement la plupart des résidus conservés au sein des AspRS jouant un rôle important dans la fixation du ligand. En particulier la Lys 198, la Gln 199, l'Asp 233, le Glu 235 sont parfaitement conservés au sein de nos séquences. En outre, l'analyse

structurale des modèles reconstruits révèle une bonne prédiction des rotamères de ces résidus. Par exemple, la lysine 198 est décrite par 49 rotamères ; les conformations de la chaîne latérale dans la structure native et mutée se superposent parfaitement (figure 4.17). Par ailleurs, nos résultats révèlent un couplage entre la Lys 198 et l'Asp 233. En effet, lorsque la Lys 198 est absente de nos prédictions, l'Asp 233 n'y apparaît pas non plus[1]. Dans les structures cristallographiques d'AspRS, cet aspartate interagit avec la Lys 198 par l'intermédiaire d'un pont salin [26], [169], [154]. Il est intéressant de voir que nos prédictions en rendent compte.

Le Glu 171 participe avec la Ser 193 et la Gln 195 à la stabilisation du groupement ammonium du ligand. Dans nos prédictions, ces trois résidus sont respectivement mutés en Asp 171[2], Asp ou Ala 193 et Asp 195. L'analyse des structures obtenues montre que les mutations S193D et Q195D préservent la stabilisation du groupement NH_3^+ du ligand. La mutation E171D demeure plus difficile à expliquer. Dans ce cas, l'Asp 171 interagit avec l'Arg 217 dont l'orientation a été légèrement modifiée. Néanmoins, ces trois mutations ne semblent pas compromettre la stabilisation du groupement ammonium du ligand.

D'autre part, le caractère aromatique de la position 229 est conservé au sein de nos prédictions et permet ainsi de garantir la stabilisation de l'adénosine de l'AMP par l'intermédiaire d'interactions van der Waals.

Résidus hydrophobes et de petite taille : Les résidus hydrophobes et de petite taille sont généralement bien prédits. On prédit ainsi des thréonines en position 169 ; contrairement aux séquences obtenues sans ajustement des énergies de référence. Les résidus Ile 232, Val 234 et Thr 452 sont eux aussi très bien prédits par notre algorithme.

[1] Ce phénomène est observé sur deux séquences uniquement dans l'alignement présenté mais apparaît aussi dans d'autres séquences non montrées ici

[2] certaines séquences conservent le glutamate à cette position

Résidus aromatiques : La prédiction des résidus aromatiques est très satisfaisante. Tous ces résidus conservent leur caractère aromatique exceptée l'His 448[3]. On relève cependant, une fréquence de tryptophanes supérieure à leur fréquence naturelle. Comme nous l'avons mentionné, ces derniers établissent des interactions van der Waals très favorables avec les autres résidus hydrophobes environnants ce qui pourrait expliquer leur sur-représentation.

Positions mal prédites : Quatre positions sur 40 demeurent toutefois mal prédites. La Ser 453 est systématiquement mutée en histidine. Sa chaîne latérale est exposée au solvant ; elle n'est donc pas soumise à des contraintes stériques. Dans la structure cristallographique d'AspRS d'*E. coli*, cette serine n'engage pas d'interactions avec les résidus environnants mais interagit avec plusieurs molécules d'eau clairement définies ce qui rend sa prédiction difficile.

Dans les séquences naturelles d'AspRS, l'Asn 472 n'est pas du tout conservée. Ceci explique donc le fait de ne pas la retrouver dans nos prédictions. Le cas de la mutation S487H/D demeure difficile à expliquer. Cette position se trouve à proximité de l'Arg 489 ce qui favorise probablement la présence d'un aspartate en position 487. Néanmoins, nous n'expliquons pas l'apparition d'une histidine à cette position.

Enfin, l'His 448 conservée chez les procaryotes, est mal reproduite par nos calculs. Dans la structure cristallographique d'AspRS de *T. thermophilus*, l'His 448 interagit avec l'od1 du ligand [169]. Dans nos prédictions elle est généralement remplacée par un glutamate qui interagit avec l'ion Mg-1. L'analyse des structures obtenues révèle une légère modification de l'orientation des atomes od1 et od2 de l'aspartate. Dorénavant, l'od1 du ligand interagit avec la Lys 198 et ne peut donc plus engager d'interactions avec l'His 448 (figure 4.17). Cela pourrait donc expliquer le fait que cette dernière ne soit pas conservée dans nos prédictions.

[3]Ce cas sera discuté au paragraphe suivant

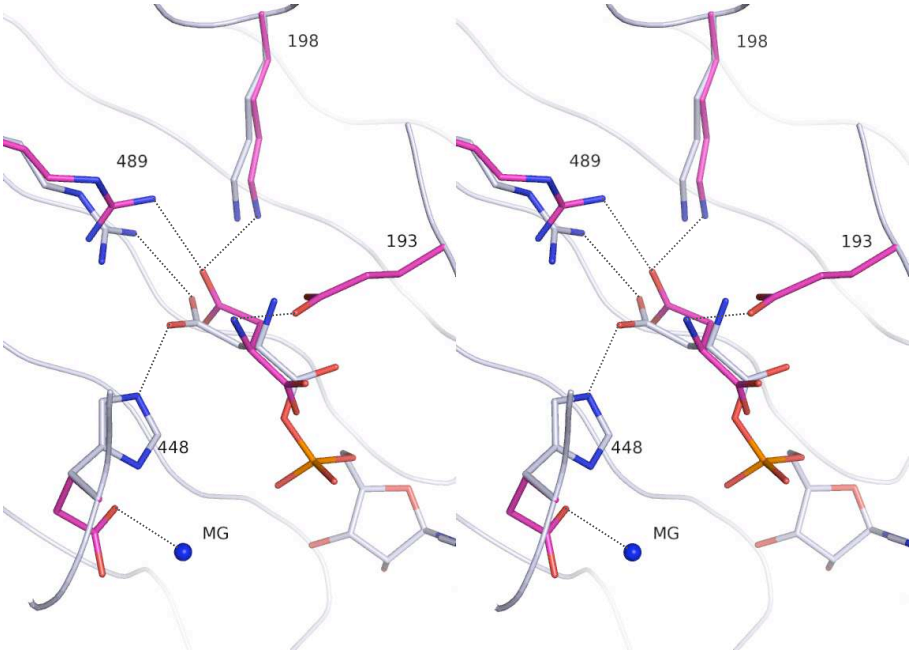


FIG. 4.17 – **Vue stéréo du site actif de l'AspRS d'*E. coli* (gris).** Le mutant est représenté en magenta. La conformation de la chaîne latérale du ligand chez le mutant est légèrement modifiée par rapport à la conformation native. Le ligand ne pourrait plus interagir avec une éventuelle histidine en position 448. Chez le mutant, l'His 448 est donc remplacée par un aspartate qui interagit avec l'ion Mg-1. Par ailleurs nous soulignons la similarité conformationnelle entre la Lys 198 issue de la structure cristallographique et de nos prédictions. Les liaisons hydrogènes et les ponts salins sont matérialisés par des lignes pointillées.

En résumé, nous constatons avec satisfaction que la plupart des résidus directement impliqués dans la liaison du ligand sont bien prédits. Bien que certains résidus importants soient mutés dans nos prédictions, les mutations semblent préserver le rôle fonctionnel des résidus concernés. Toutefois, il serait nécessaire de réaliser des tests expérimentaux pour se prononcer clairement. Notre procédure propose des séquences présentant une certaine diversité tout en maintenant les prérequis structuraux nécessaires au bon fonctionnement de la protéine. Il est alors intéressant de voir l'influence du ligand sur les séquences produites par notre algorithme, en particulier, les séquences obtenues en remplaçant l'AspAMP par l'AsnAMP.

Analyse des jeux de séquences de meilleur score BLOSUM avec l'AsnAMP

Une nouvelle matrice d'énergie fut calculée en présence de l'AsnAMP. Comme précédemment, les séquences obtenues furent classées selon leur score BLOSUM. Ces séquences étant prédites en présence de l'AsnAMP dans le site actif, on s'attend à observer la mutation des résidus caractéristiques de l'AspRS et à obtenir des séquences présentant des similitudes avec l'AsnRS. Il n'aurait donc pas été pertinent de comparer nos séquences à la séquence native d'AspRS pour le calcul du score BLOSUM. Ainsi, nous avons défini une séquence consensus à partir d'un alignement structural de plusieurs structures d'Asprs et d'AsnRS [105]. L'alignement fut effectué à l'aide de MATRAS. Il permit de définir une séquence consensus pour les différentes séquences d'AsnRS alignées. Cette séquence consensus fut ensuite utilisée comme cible pour le calcul du score BLOSUM. Les 40 positions actives de cette séquence consensus sont présentées dans la figure 4.19, tandis que le protocole nécessaire pour leur obtention est présenté en figure 4.18.

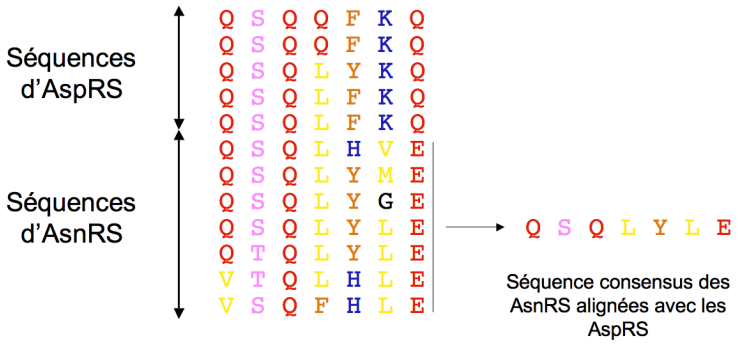


FIG. 4.18 – **Obtention de la séquence consensus à partir d'un alignement structural d'AspRS et d'AsnRS.** Une fois les différentes synthétases alignées, on détermine le consensus des séquences d'AsnRS.

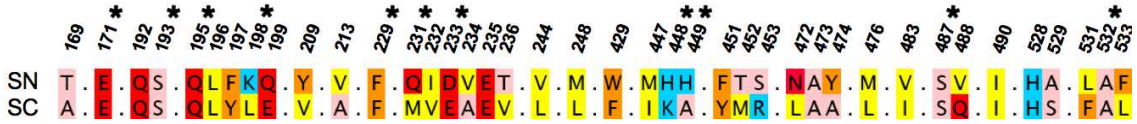


FIG. 4.19 – **Représentation des 40 positions 'Actives' extraites de la séquence native d'AspRS d'*E. coli* (SN) ou de la séquence consensus d'AsnRS (SC).**

Les meilleures séquences prédites par Proteus présentent un score de 70 tandis que la séquence consensus contre elle même donne un score de 187. Les identités de séquence avec la séquence consensus atteignant 37%. Les 25 meilleures séquences sont présentées en figure 4.20. L'accord entre les séquences prédites et la séquence consensus est excellent. Généralement, les propriétés (hydrophobes, aromatiques, polaires ou chargés) des résidus de la séquence consensus sont conservées. Le résultat le plus remarquable est l'absence de lysine en position 198. Dans nos prédictions, cette position est occupée par une leucine présente aussi dans de nombreuses séquences naturelles d'AsnRS. Notre procédure a donc réussi à capturer le rôle fonctionnel de cette lysine, puisqu'elle n'apparaît plus en présence du ligand AsnAMP.

En outre les acides aminés de petite taille et les hydrophobes (aromatiques inclus) de la séquence consensus sont très bien prédits par notre algorithme. Comme précédemment, on relève parfois le remplacement de certains résidus hydrophobes non aromatiques par des aromatiques. Ceci est visiblement dû à notre fonction d'énergie qui favorise l'enfouissement des résidus hydrophobes aromatiques. Ces cas ont largement été expliqués dans la section précédente.

Trois positions demeurent mal prédites. Nous ne reviendrons pas sur le cas de la position 453 exposée au solvant (conférer la section précédente). En position 233, nous prédisons un mélange d'asparagines, thréonines, serines et valines. Le rôle fonctionnel de l'aspartate conservé chez toutes les AspRS, ou du glutamate spécifique des AsnRS n'est donc pas retrouvé.

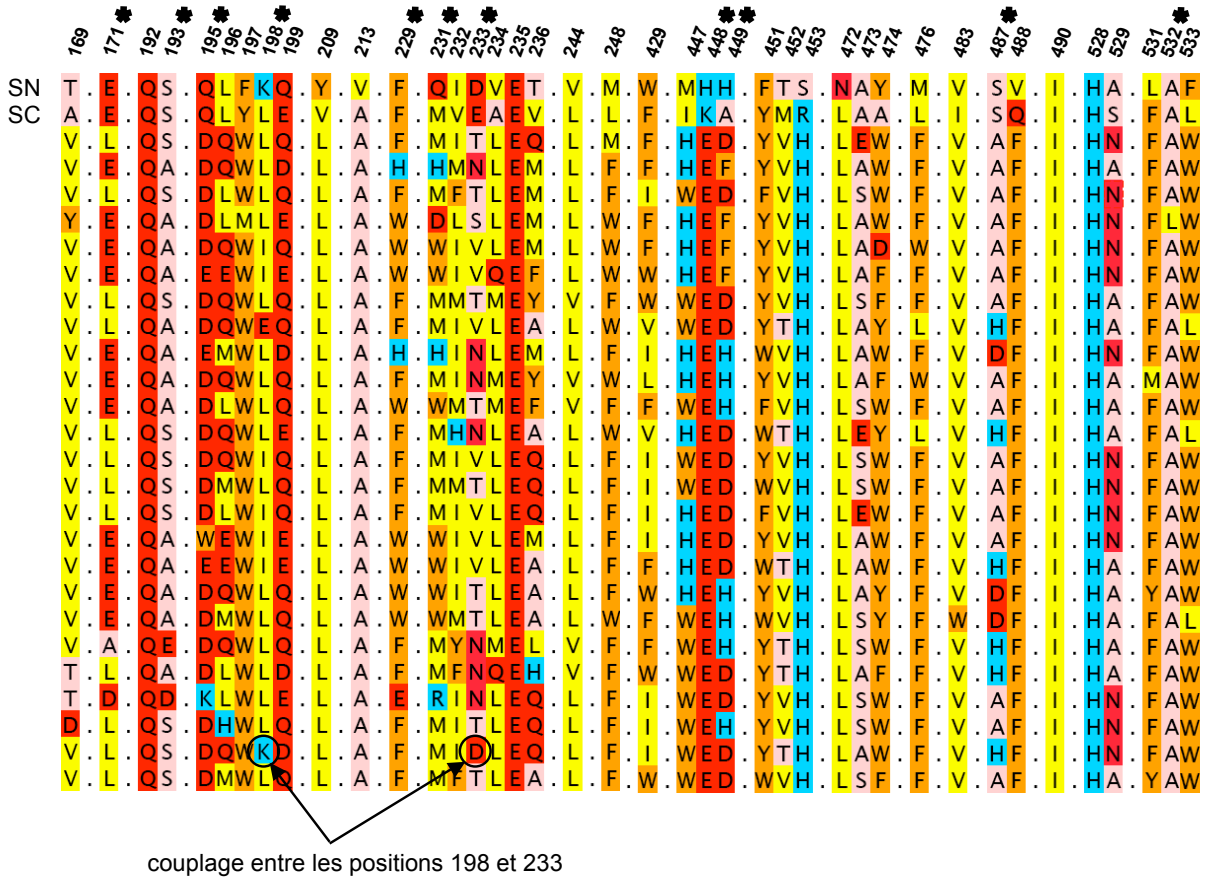


FIG. 4.20 – **Alignement de positions des séquences de meilleurs score BLOSUM obtenues en présence de l'AsnAMP.** Pour plus de clarté, seules les 40 positions actives sont représentées, les autres positions demeurant invariantes. 'SN' indique la séquence native d'AspRS tandis que 'SC' indique la séquence consensus d'AsnRS. Les résidus à proximité du ligand sont marqués par une étoile.

La position 448 est occupée par une histidine chez les AspRS de procaryotes et par une lysine chez toutes les AsnRS. Dans la structure cristallographique de l'AsnRS de *T. thermophilus*, cette lysine interagit avec l'Asp 352 [475] également impliqué dans un pont salin avec l'ion Mg-1 (numérotation de *T. thermophilus* et

[d'AspRS d'*E. coli*]). Or la boucle sur laquelle elle se trouve ne présente pas la même conformation que dans la structure de l'AspRS d'*E. coli* utilisée pour nos calculs (figures 4.21). Ainsi, une lysine en position 448 ne pourrait pas interagir avec l'Asp 475 présent chez l'AspRS d'*E. coli* et ne serait donc pas stabilisée.

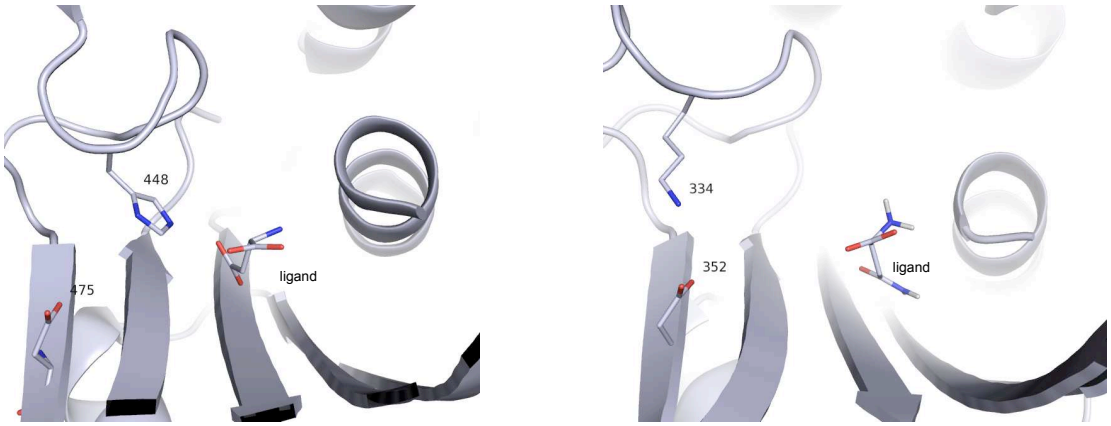


FIG. 4.21 – **A gauche :** le site actif de l'AspRS d'*E. coli*. Une lysine en position 448, ne pourrait pas interagir avec l'Asp 475, la boucle flexible n'ayant pas exactement la même conformation dans l'AspRS et l'AsnRS. **A droite :** le site actif de l'AsnRS de *T. thermophilus*. La Lys 334 et l'Asp 352 correspondant respectivement aux positions 448 et 475 de l'AspRS interagissent par l'intermédiaire d'un pont salin.

Nous avons donc montré que notre procédure de CPD était sensible à la nature du ligand dans la poche. En introduisant une contrainte biologique (nature du ligand) nous arrivons à obtenir des séquences présentant des caractéristiques structurales mais aussi fonctionnelles de l'AspRS. En effet, nous constatons qu'en fonction du ligand présent dans la poche nous retrouvons ou non la lysine 198 qui joue un rôle déterminant dans la fonction de l'enzyme. Ce résultat est encourageant pour l'ingénierie d'une synthétase de spécificité modifiée.

4.2.2 Prédiction de changements d'affinité associés à des mutations ponctuelles

Nous avons également évalué la capacité de notre procédure de CPD à reproduire des changements d'affinité associés à des mutations ponctuelles pour lesquelles des données expérimentales étaient disponibles [26]. En effet, les interactions protéine-ligand sont très importantes dans l'optique d'une ingénierie de la spécificité de l'AsnRS. Les détails sont donnés dans l'article de Schmidt et coll. en section 3.2. Sur l'ensemble des neuf mutations testées, nous obtenons une erreur moyenne de 1.89 kcal/mol (table 4.4).

		Protocole 'minimal'			
Mutations	$\Delta\Delta$G (EXP)	ΔG_{cplx} (NAT)	ΔG_{cplx} (MUT)	$\Delta\Delta$G (PM)	Erreur
Q195A	1.59	-65.12	-62.94	2.18	0.59
Q195E	1.69	-65.12	-64.15	0.97	0.72
Q195N	2.27	-65.12	-65.97	-0.85	3.12
Q199A	1.68	-65.12	-65.08	0.04	1.64
Q199E	0.70	-65.12	-65.28	-0.16	0.86
Q199N	1.74	-65.12	-65.05	0.07	1.67
E235D	1.06	-65.12	-65.36	-0.24	1.30
R489H	1.35	-65.12	-60.52	4.61	3.26
Moyenne					1.89

Table 4.4. Erreur moyenne des changements d'affinité associés à des mutations ponctuelles (Protocole 'minimal'). Les résultats sont donnés en kcal/mol. ΔG_{cplx} correspond à l'énergie de liaison du complexe natif (NAT) et du complexe mutant (MUT). $\Delta\Delta$G correspond à la différence d'énergie libre de liaison entre le complexe natif et mutant dans les conditions expérimentales (EXP) ou calculée avec le protocole 'minimal' (PM). Les calculs ont été réalisés avec : $\epsilon = 16$, modèle de solvant = CASA/PHIA, α=0.5.

Le protocole 'Proteus' fut ensuite testé à son tour. Dans ce cas, toutes les chaînes latérales sont autorisées à bouger contrairement au protocole 'minimal' où seule la chaîne latérale du résidu muté varie. Pour ce faire, nous sommes repartis de la matrice d'énergie calculée précédemment. Proteus fut lancé en maintenant fixée la séquence (incluant la mutation désirée) afin de déterminer le jeu de rotamères optimal. Les résultats obtenus avec ce protocole sont présentés en table 4.5.

	Protocole 'Proteus'				
Mutations	$\Delta\Delta$G (EXP)	ΔG_{cplx} (NAT)	ΔG_{cplx} (MUT)	$\Delta\Delta$G (PP)	Err. moyenne
Q195A	1.59	-26.37 ± 8	-25.34 ± 6	1.03	0.56
Q195E	1.69	-26.37 ± 8	-31.00 ± 3	-4.63	6.32
Q195N	2.27	-26.37 ± 8	-24.90 ± 8	1.47	0.80
Q199A	1.68	-26.37 ± 8	-26.15 ± 8	0.22	1.46
Q199E	0.70	-26.37 ± 8	-29.34 ± 2	-2.97	3.67
Q199N	1.74	-26.37 ± 8	-26.26 ± 8	0.11	1.63
D233E	1.68	-26.37 ± 8	-26.69 ± 9	-0.32	2.00
E235D	1.06	-26.37 ± 8	-26.28 ± 8	0.09	0.97
R489H	1.35	-26.37 ± 8	-24.49 ± 9	1.88	0.53
Moyenne					1.99

Table 4.5 : Erreur moyenne des changements d'affinité associés à des mutations ponctuelles (Protocole 'Proteus'). Les résultats sont donnés en kcal/mol. ΔG_{cplx} correspond à l'énergie de liaison du complexe natif (NAT) et mutant (MUT). $\Delta\Delta$G correspond à la différence d'énergie libre de liaison entre le complexe natif et mutant dans les conditions expérimentales (EXP) ou calculée avec le protocole 'Proteus' (PP). Les calculs ont été réalisés avec : $\epsilon = 16$, modèle de solvant = CASA/PHIA, α=0.5. Pour chaque mutant, l'énergie de liaison, ΔG_{cplx}, correspond à une moyenne des énergies de liaisons obtenues pour un ensemble de 1000 conformations de basse énergie.

Nous observons une erreur moyenne de 1.99 kcal/mol ce qui paraît satisfaisant. Dans ce nouveau protocole le nombre de degrés de liberté est beaucoup plus grand que dans le protocole 'minimal'. Nous rappelons que notre système comprend environ 250 résidus (Gly, Pro et Cys exclues), et que les mutations étudiées impliquent des résidus polaires ou chargés du site actif. Cependant il faut noter que l'écart type des ΔG_{cplx} est relativement important. Ceci suggère au moins deux modes d'associations du complexe. Cette observation sera détaillée au chapitre suivant. Par ailleurs dans le cas du protocole 'Proteus', nous avons pu estimer le $\Delta\Delta G$ associé à la mutation D233E. Seuls les atomes du squelette peptidique sont fixés, ainsi, la molécule d'eau et la chaîne latérale du Glu 233 ont pu se réajuster pour éviter le conflit stérique observé dans le protocole 'minimal'.

Chapitre 5

Evolution dirigée de l'AsnRS

Dans le chapitre précédent nous avons évalué notre procédure de CPD dans le *design* du site actif de l'AspRS et montré que nous étions capables de produire des séquences présentant des caractéristiques structurales et fonctionnelles de l'AspRS. Dans ce chapitre nous appliquons notre procédure au site actif de l'AsnRS qui présente de fortes similitudes avec celui de l'AspRS. Une première étape consista à estimer les énergies de liaison entre l'AsnRS native de *T. thermophilus* et les deux ligands (AsnAMP et AspAMP) en utilisant notre modèle énergétique et notre protocole de 'reconstruction' des modèles.

5.1 Estimation des affinité de l'AsnRS native pour l'AsnAMP et l'AspAMP

Le paysage énergétique décrit par nos champs de force ne représente pas le 'vrai' paysage énergétique. Par conséquent, nous avons pris pour parti de caractériser une protéine par un ensemble de conformations de basse énergie plutôt que de chercher à déterminer 'la' conformation de plus basse énergie qui ne correspondrait probablement pas à la conformation native. Ainsi, la séquence native étant donnée, nous avons procédé à l'exploration de son espace conformationnel afin de déterminer cet

ensemble de conformations de basse énergie. Les 2000 jeux de rotamères de plus basse énergie furent retenus et reconstruits par notre protocole de 'reconstruction'. Les énergies de liaison des complexes AsnRS :AsnAMP et AsnRS :AspAMP furent obtenues en calculant la moyenne des énergies de liaison des 2000 solutions retenues.

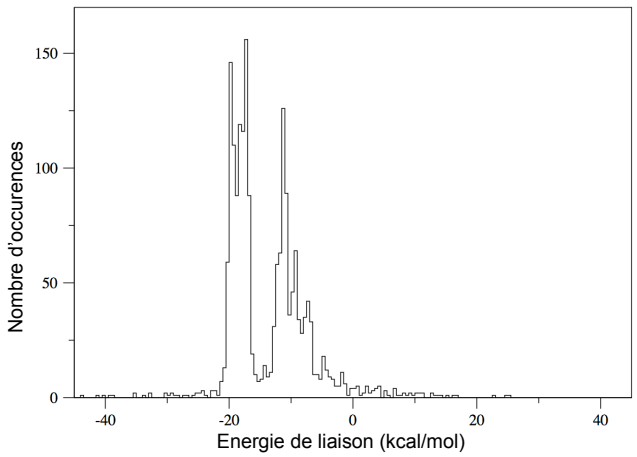


FIG. 5.1 – **Distribution des énergies de liaison des complexes AsnRS :AsnAMP pour les 2000 modèles reconstruits.**

Les résultats pour l'AsnAMP sont présentés figure 5.1. Nous observons deux pics distincts correspondant à au moins deux modes d'association différents. Le premier et le second pic correspondent respectivement à des énergies de liaison moyennes de -19.3 kcal/mol et -8.4 kcal/mol. Nous avons analysé la variation des contributions énergétiques du ligand et des différents résidus du site actif dans l'énergie globale de la protéine (figure 5.2). Le ligand, le Glu 164, la Gln 187 et la Met 338 présentent des écarts type supérieurs aux autres. Ce résultat suggère un changement conformationnel de ces derniers. Il est intéressant de remarquer que les résidus présentant les plus grandes variations d'énergie au sein des différents modèles sont généralement en interaction directe avec le ligand. Nous avons tenté de caractériser ces deux populations d'un point de vue conformationnel.

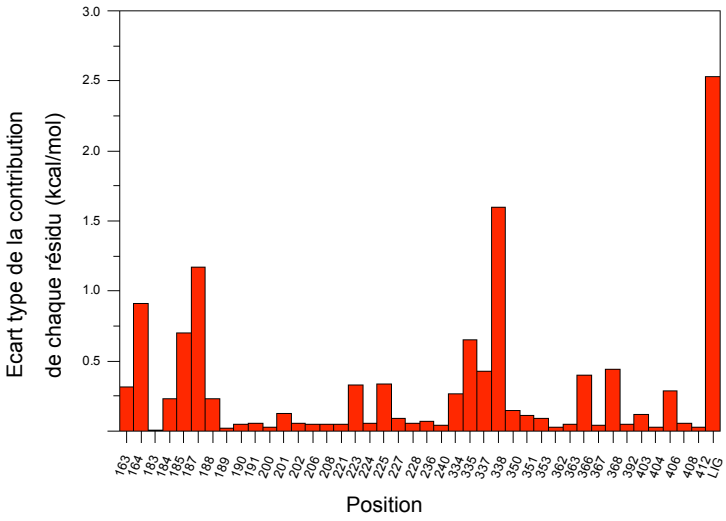


FIG. 5.2 – **Représentation de l'écart type de la contribution énergétique des principaux résidus du site actif.** La contribution énergétique est calculée sur l'ensemble des 2000 jeux de rotamères.

Nous avons donc comparé les structures peuplant chacun des pics à la structure cristallographique. Nous calculons alors un RMSD moyen de 1.87 ±0.12 Å entre le site actif de la structure cristallographique et l'ensemble des conformations du premier pic et un RMSD moyen de 2.29 ±0.25 Å avec le second pic.

Nous avons précisé l'étude en nous intéressant à la contribution du ligand et des différents résidus du site actif dans le RMSD global. On a constaté que le ligand subit des changements majeurs de conformations. La figure 5.3 présente pour chacun des deux pics, la distribution des angles dièdres du ligand et des principaux résidus de la poche. On s'aperçoit que pour les deux pics, le résidu 164 échantillonne différents conformères. Le Glu 164, présent sur une des boucles flexibles du site actif, est exposé au solvant et n'est donc pas soumis à des contraintes stériques. Dans le cas du premier pic, les chaînes latérales des autres résidus sont représentées par une conformation unique proche de la conformation native. Ce phénomène est observé également dans les structures peuplant le second pic exceptés les résidus 187 et 337. Par ailleurs, dans les structures du premier pic, l'analyse des angles dièdres du ligand révèle que ce dernier est décrit par un rotamère unique similaire

à l'orientation observée dans la structure cristallographique. A l'inverse, dans le cas du second pic ; le ligand échantillonne différentes conformations.

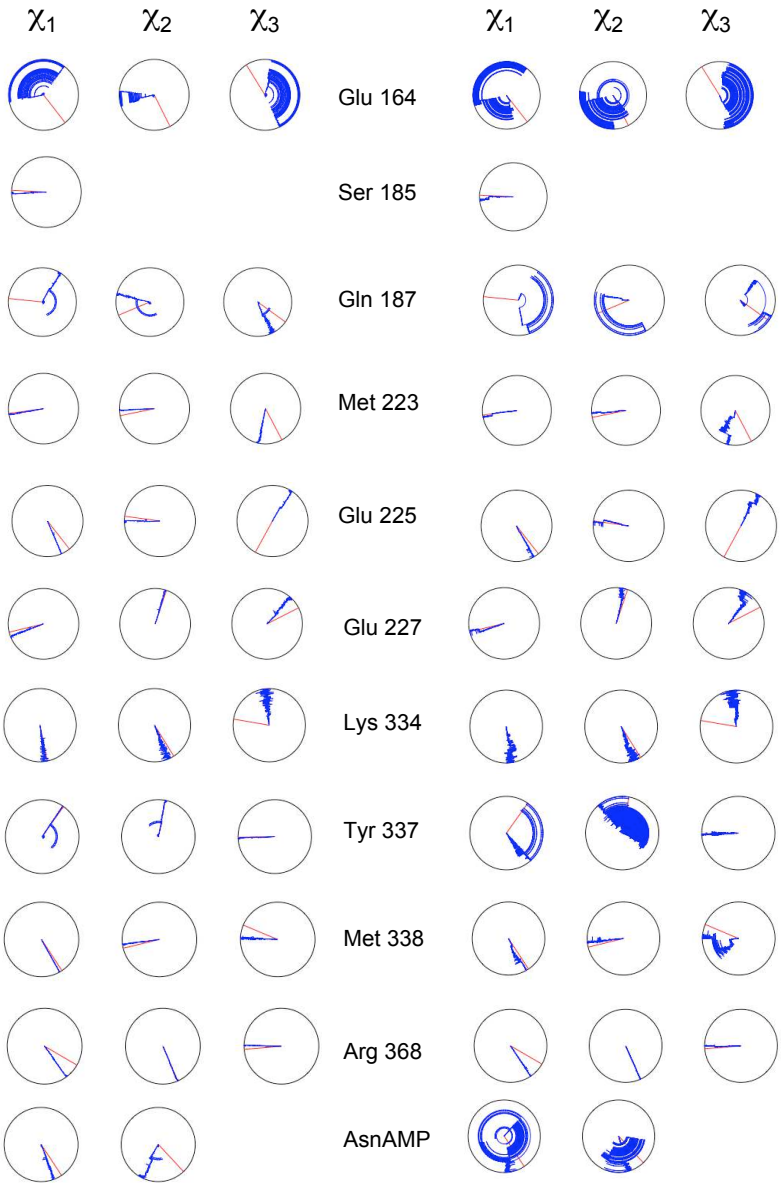


FIG. 5.3 – **Représentation des angles dièdres χ_1, χ_2, χ_3 du ligand et des résidus de la poche à proximité de celui-ci. A gauche :** les conformères peuplant le premier pic. **A droite :** ceux peuplant le second pic. L'angle dièdre natif est indiqué en rouge.

La figure 5.4 représente un ensemble de conformations du ligand peuplant le premier pic. Ces dernières présentent une orientation proche de l'orientation native. Le RMSD calculé entre la conformation moyenne du ligand du pic 1 et celle issue de la structure cristallographique est de 1.3 Å La déviation structurale maximale entre deux structures au sein du pic est de 0.21 Å. L'interaction entre l'od1 du ligand et le groupement ammonium de l'Arg 368 est maintenue. De même, le groupement NH_2 de l'asparagine conserve ses interactions favorables avec les résidus 225 et 227.

Nous avons regroupé les conformations du ligand peuplant le second pic par similarité structurale. On identifie trois conformations majoritaires peuplées respectivement par 156, 495 et 207 conformères (figure 5.4). La déviation structurale au sein de chacun des sous-groupes est respectivement de 0.02 Å, 0.17 Å, et 0.4 Å. Les conformères peuplant les deux premiers, sont bien distincts de la conformation native. Leurs structures moyennes respectives présentent un RMSD avec la structure native de 2.9 et 3.8 Å. Dans ces deux sous-groupes, le groupement NH_2 de la chaîne latérale du ligand n'interagit plus avec les résidus 225 et 227 expliquant probablement la baisse d'affinité observée pour le pic 2. Cependant l'od1 de l'asparagine semble tout de même interagir avec l'Arg 368 ce qui pourrait aussi expliquer le fait que l'on obtienne, pour l'ensemble des structures peuplant le second pic, une énergie de liaison favorable malgré l'orientation non native du ligand. Les conformères du troisième sous-groupe présentent une orientation similaire à celles du premier pic. Néanmoins, dans les structures de ce sous-groupe l'axe Cβ-Cγ subit une légère rotation par rapport à celui des conformations du premier pic. Ainsi, l'od1 du ligand se trouve plus éloigné du groupement positif de l'Arg 368 (2.9 Å) que dans la structure caractéristique du pic 1 (2.6 Å). De même, le groupement NH_2 se trouve à 4.3 Å du groupement carboxylate du Glu 225 dans la conformation du premier pic tandis que plus de 5.4 Å les séparent dans les structures peuplant le troisième sous-groupe. Ceci expliquerait peut être la différence d'affinité observée.

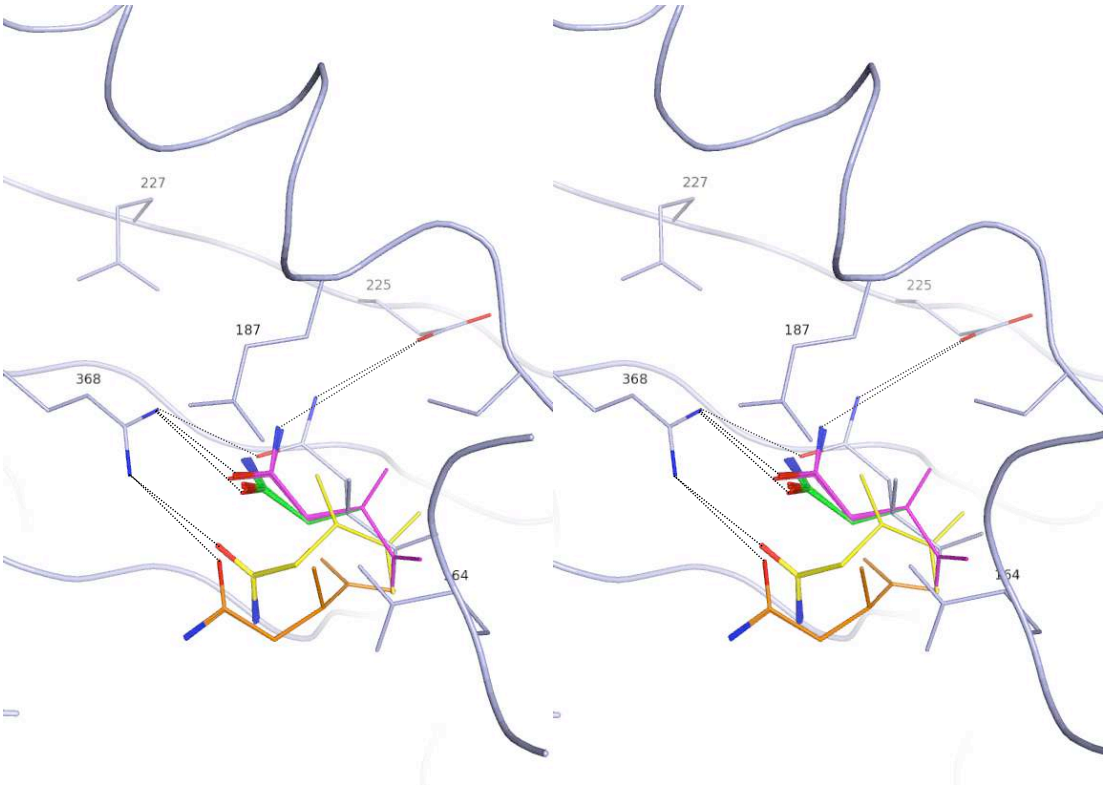


FIG. 5.4 – **Représentation d'un échantillon de conformères caractéristiques des deux pics.** En gris le site actif de la structure native d'AsnRS de *T. thermophilus* et le ligand natif. En magenta : échantillon de conformations du premier pic. En vert : second pic, premier sous-groupe. En jaune : second pic, deuxième sous-groupe. En orange : second pic, troisième sous-groupe. Les principaux résidus de la poche sont représentés en 'sticks'. Seuls les atomes impliqués dans les interactions d'intérêt sont colorés selon leur type atomique (azote en bleu et oxygène en rouge). Les liaisons hydrogènes et ponts salins sont matérialisées par des lignes pointillées.

Pour la suite de nos travaux nous considérerons les conformations peuplant le pic 1 comme représentatives de la l'AsnRS native. Par conséquent, nous estimerons l'énergie de liaison entre l'AsnRS native et l'asparagine aux alentours de -19.3 kcal/mol. La même étude fut réalisée pour le complexe AsnRS :AspAMP. Nous obtenons une population caractérisée par une énergie de liaison moyenne de -11.7 kcal/mol (figure 5.5). Cette population est représentée par un pic unique.

Néanmoins, les énergies de liaisons des différents complexes prédits sont très variables ; nous mesurons une déviation standard d'environ 8.5 kcal/mol. Ce phénomène est comparable à celui observé pour le système de l'AspRS (conférer section 4.2.2). L'ensemble de cette étude montre qu'une séquence donnée peut être caractérisée par un ensemble de conformères d'énergie similaire. Dans le cas du complexe Asn :AsnRS, le conformère natif fut déterminé aisément dans la mesure où nous possédions la structure cristallographique du complexe. Il suffisait de comparer l'ensemble des conformères prédits à la structure cristallographique. Cependant, dans le cas des mutants, nous ne possédons pas nécessairement d'information structurale. Ainsi, il est difficile de pondérer les différentes conformations observées et de déterminer la ou les conformation(s) native(s). Nous avons donc pris pour parti de caractériser une séquence donnée par un ensemble de conformations de basse énergie. Pour la suite de notre étude, les deux énergies calculées ci-dessus, pour les complexes Asn :AsnRS et Asp :AsnRS, joueront le rôle d'énergies 'témoins' qu'il sera intéressant de comparer aux énergies de liaison que nous obtiendrons pour les mutants.

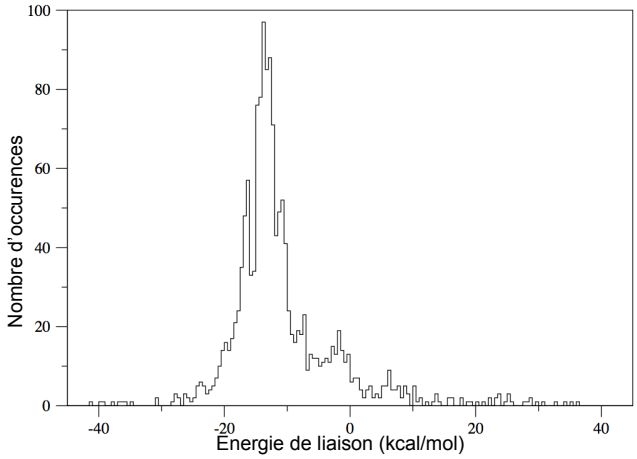


FIG. 5.5 – **Distribution des énergies de liaison des complexes AsnRS :AspAMP pour les 2000 modèles reconstruits.**

5.2 Analyse des séquences d'AsnRS obtenues en présence de l'AsnAMP

Afin de contrôler la qualité des résultats, nous avons d'abord analysé les séquences obtenues pour le système AsnRS :AsnAMP. De même que pour l'AspRS, nous avons classé les séquences produites selon leur score BLOSUM. Nous rappelons que notre but est d'isoler des mutants de stabilité raisonnable spécifiques de l'aspartate et non pas des mutants présentant une stabilité maximale. Les 25 meilleures solutions sont présentées en figure 5.6.

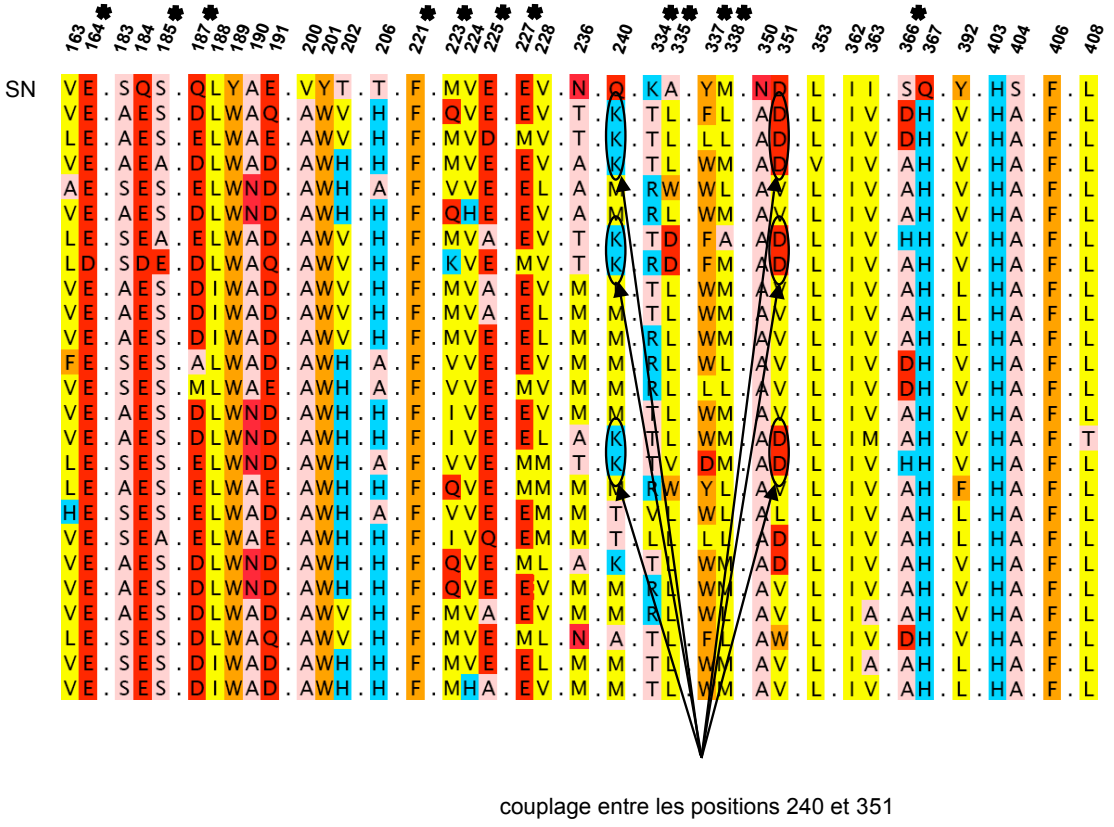


FIG. 5.6 – **Alignement de positions des séquences de meilleurs score BLOSUM obtenues en présence de l'AsnAMP.** La première ligne indique la numérotation des positions selon la séquence d'AsnRS de *T. thermophilus*. Seules les 38 positions 'Actives' sont représentées. Les résidus à proximité du ligand sont marqués d'une étoile. 'SN' indique la séquence native d'AsnRS. Le code couleur demeure inchangé.

La figure 5.7 montre l'alignement de séquences naturelles d'AsnRS. Les meilleures séquences calculées présentent des scores BLOSUM de 82 tandis que la séquence native contre elle même donne un score de 191. Le pourcentage d'identité stricte entre les séquences prédites et la native est aux alentours de 40%. De même que pour l'AspRS, les acides aminés hydrophobes (aromatiques inclus) et ceux de petite taille sont très bien prédits, exceptée les thréonines 202 et 206 , généralement remplacées par une histidine. L'Ala 335 située dans le site actif est souvent remplacée par une leucine. On suppose que cela est dû à notre fonction d'énergie qui tend à enfouir préférentiellement des acides aminés de type hydrophobes afin de maximiser la stabilité de la protéine.

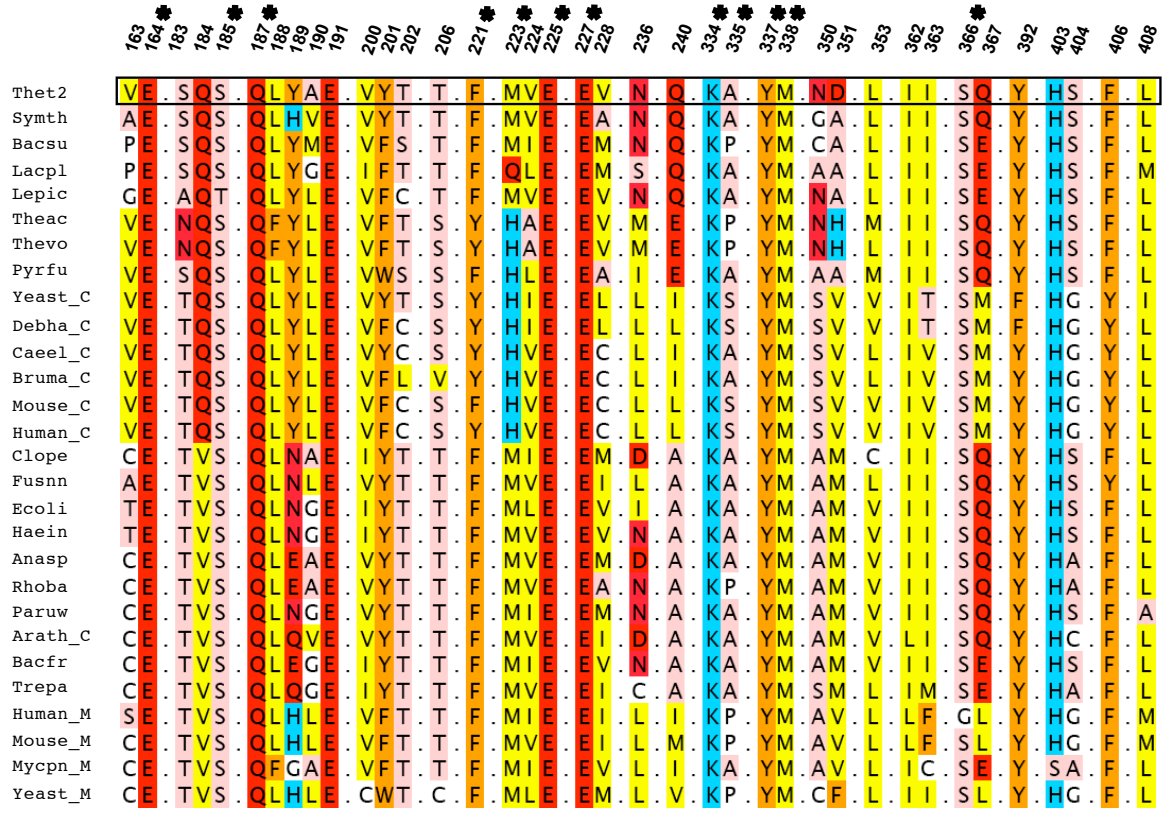


FIG. 5.7 – **Alignement de positions des séquences naturelles d'AsnRS provenant de différents organismes.** Le code couleur reste inchangé, seules les 38 positions 'Actives' sont représentées. La séquence d'AsnRS issues de *T. thermophilus* est encadrée en noir. '_C' réfère aux AsnRS cytoplasmiques tandis que '_M' réfère aux AsnRS mitochondriales.

Les résidus polaires et chargés s'avèrent, eux aussi, prédits correctement. Nous soulignerons en particulier le cas des positions 164, 191, 225 et 227 qui conservent globalement leur charge négative. En outre, nous pouvons expliquer le cas des positions 236, 240, 350 et 351, difficilement conservées au sein de nos prédictions. En effet, l'analyse de l'alignement des séquences naturelles d'AsnRS montre que ces quatre positions ne sont pas du tout conservées parmi les AsnRS. Par ailleurs nous remarquons un autre phénomène de couplage entre les positions 240 et 351. Dans la structure cristallographique d'AsnRS de *T. thermophilus*, la Gln 240 forme une liaison hydrogène avec l'Asp 351. Il est intéressant d'observer que lorsque nous prédisons l'aspartate natif en position 351, nous prédisons systématiquement une lysine en position 240. Cette dernière réalise un pont salin avec l'Asp 351 figure 5.8.

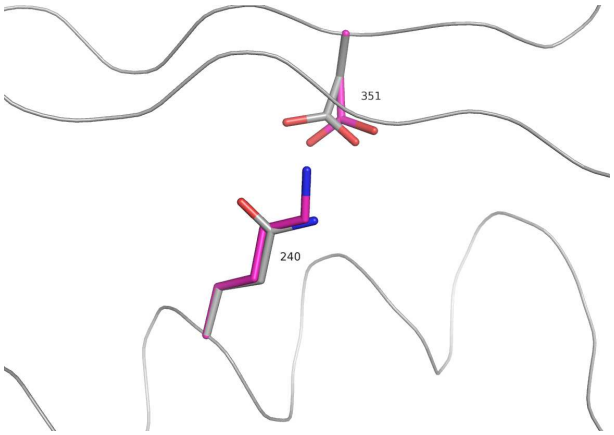


FIG. 5.8 – **Interaction entre les résidus 240 et 351.** La structure native est représentée en gris tandis que la mutante est colorée en magenta.

Quatre positions demeurent mal prédites. La Gln 184 est systématiquement remplacée par un glutamate. L'analyse structurale des modèles révèle un pont salin entre ce glutamate et l'Arg 208 dont l'orientation a été légèrement modifiée. Le même phénomène avait été observé pour l'ApsRS d'*E. coli* en section 4.2.1. De même, la Gln 187 conservée chez toutes les AsnRS, se trouve mutée en aspartate ou

en glutamate. Cette dernière participe à la stabilisation du groupement ammonium de l'asparagine par l'intermédiaire d'une liaison hydrogène [12]. L'analyse structurale des mutants obtenus montre qu'un aspartate ou un glutamate préserve en cette position le rôle de la Gln 187 en réalisant un pont salin avec le groupement ammonium du ligand. Le cas de la Gln 367 est similaire à celui de la Ser 453 chez l'AspRS décrit en section 4.2.1. Cette position est exposée au solvant et n'est pas conservée au sein des AsnRS naturelles ni au sein de nos prédictions. Nous n'expliquons pas le cas de la Lys 334. Cette dernière, nous l'avons vu en section 4.2.1., interagit avec l'Asp 352 conservé chez toutes les AsnRS. Cette interaction semble préservée par la mutation K334R dans certaines de nos prédictions. Néanmoins, dans la plupart des cas cette lysine est remplacée par une thréonine ne pouvant pas interagir avec l'Asp 352. Cette étape nous a permis de contrôler la qualité de nos prédictions et ainsi être en mesure d'aborder l'étape d'évolution dirigée de l'AsnRS.

5.3 Analyse des séquences d'AsnRS obtenues en présence de l'AspAMP

Dans cette section, nous analysons les mutants obtenus en présence d'AspAMP dans le site actif. De même que pour l'AspRS, nous avons défini une séquence consensus. Le protocole nécessaire pour l'obtention de cette séquence est décrit en section 4.2.1. Cette séquence correspond à la séquence consensus des AspRS alignées structuralement avec différentes AsnRS et sera utilisée comme cible pour le calcul du score BLOSUM. Les 38 positions 'Actives' de la séquence native d'AsnRS et de la séquence consensus sont présentées dans la figure 5.9. Nous avons classé les séquences par leur score BLOSUM et récupéré les 2000 meilleures. Ces dernières furent ensuite reconstruites en 3D par notre protocole de reconstruction. Ces 2000 mutants présentent

des énergies comparables. La différence entre ceux présentant les moins bonnes et les meilleures énergies n'excède pas 15 kcal/mol.

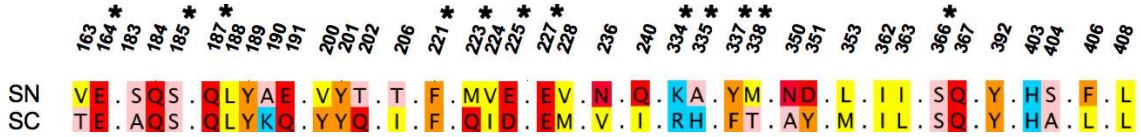


FIG. 5.9 – **Représentation des 38 positions 'Actives' extraites de la séquence native d'AsnRS de *T. thermophilus* (SN) ou de la séquence consensus d'AspRS (SC).**

Pour chacun des 2000 mutants reconstruit nous avons calculé l'énergie de liaison du complexe AsnRS :AspAMP. Les mutants furent classés selon leur affinité pour l'AspAMP et les 100 meilleurs furent retenus. Ces 2000 mutants présentent des énergies similaires et notre but est de maximiser l'affinité des mutants pour l'AspAMP et non de maximiser leur stabilité. Par ailleurs, nous avons pris pour parti de caractériser une protéine donnée par un ensemble de conformations de basse énergies. Aussi, pour chacun des 100 meilleurs mutants nous avons procédé à une nouvelle étape d'optimisation des rotamères avec Proteus en maintenant la séquence fixée. Un total de 20000 cycles heuristiques est réalisé pour chaque mutant. Après avoir éliminé les jeux de rotamères redondants, les 500 jeux de plus basse énergie sont retenus pour représenter chaque mutant. Ces 500 jeux sont reconstruits en 3D et on détermine l'énergie de liaison entre chacun des 500 modèles et l'aspartate. Finalement les 100 mutants de départ sont chacun caractérisés par une énergie de liaison moyenne calculée sur cet ensemble de 500 conformations de basse énergie. En parallèle, la procédure est répétée pour chacun des 100 mutants en présence de l'AsnAMP cette fois afin d'identifier les mutants présentant une meilleure affinité pour l'AspAMP que pour l'AsnAMP. L'ensemble de la procédure est schématisé par la figure 5.10.

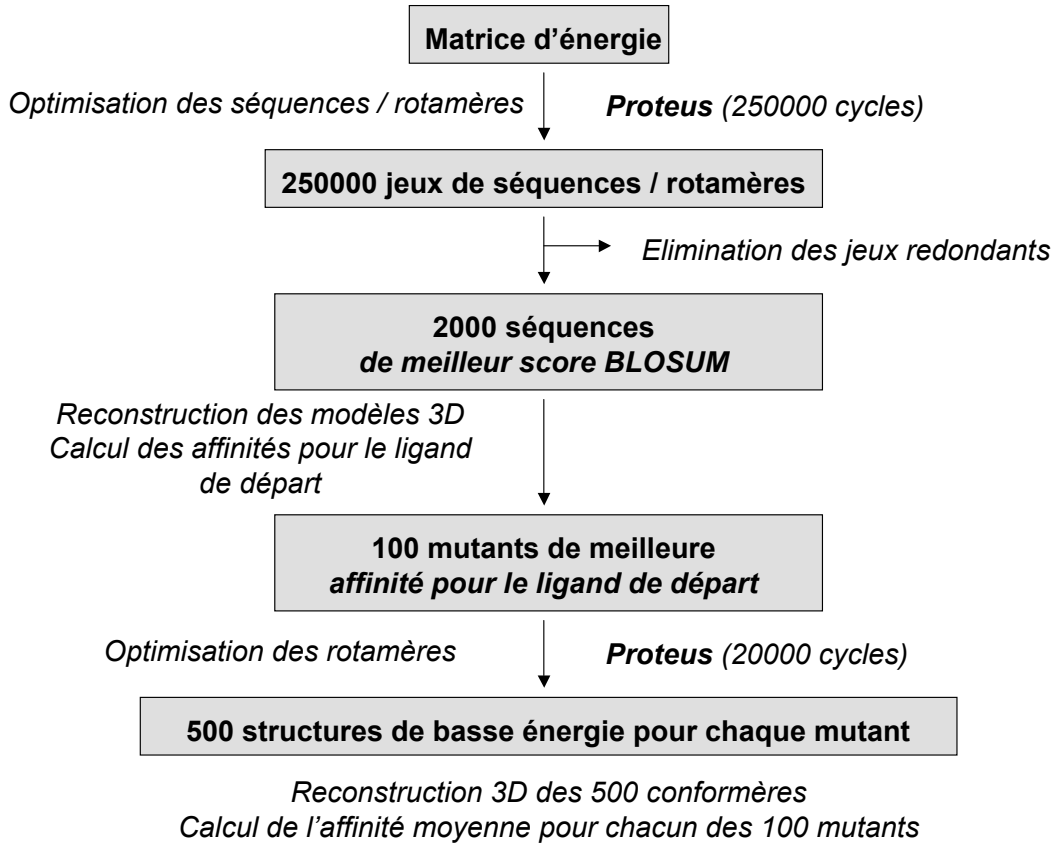


FIG. 5.10 – **Schéma représentant l'ensemble de la procédure en partant de la génération de la matrice d'énergie jusqu'au calcul d'affinité moyenne caractérisant chacun des 100 mutants les plus prometteurs.**

Cette procédure nous permit d'identifier 86 mutants présentant une meilleure affinité pour l'AspAMP que pour l'AsnAMP. Les séquences des 30 meilleurs mutants sont présentées en figure 5.11. Nous constatons qu'en mutant le ligand natif en AspAMP, nous avons favorisé l'introduction de nombreuses charges dans le site actif.

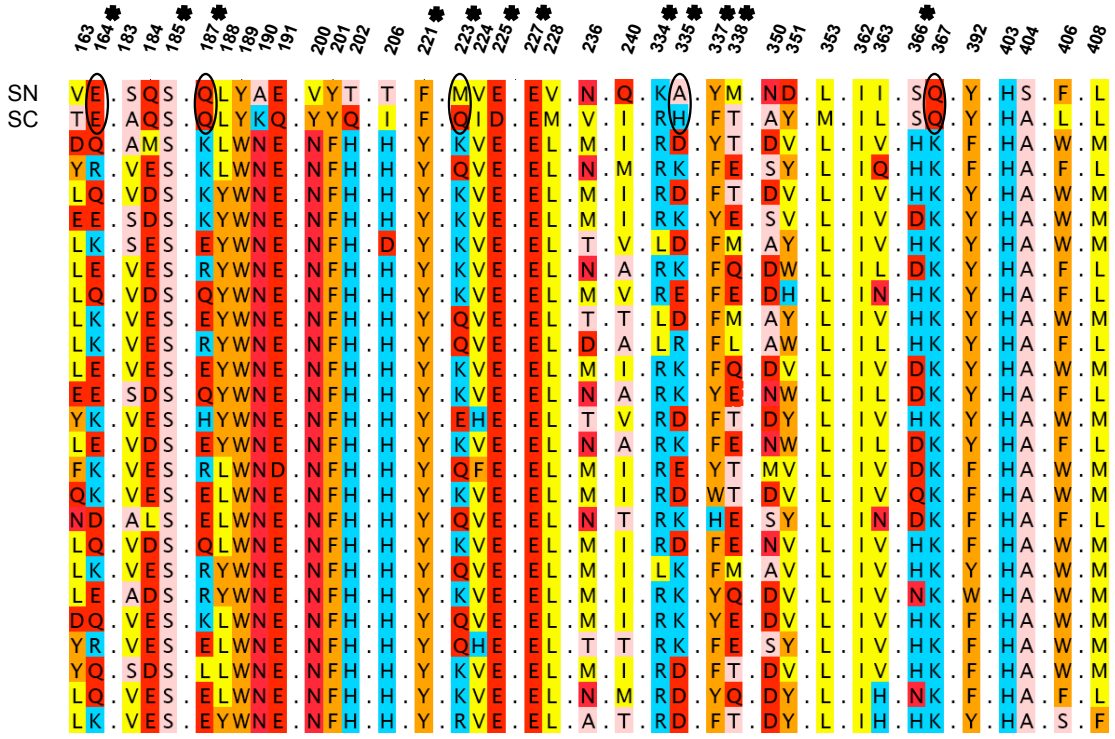


FIG. 5.11 – **Alignement de position des séquences de meilleurs score BLOSUM obtenues en présence de l'AspAMP.** Pour plus de clarté, seules les 38 positions 'Actives' sont représentées, les autres positions demeurant invariantes. La première ligne 'SN' représente la séquence native d'AsnRS de *T. thermophylus* tandis que la ligne suivante 'SC' correspond à la séquence consensus d'AspRS. Les cinq positions pour lesquelles nous observons l'introduction de charges positives sont entourées en noir.

La position 190 correspond à la position 198 de l'AspRS d'*E. coli*, occupée systématiquement par une lysine participant à la spécificité de l'interaction AspRS :AspAMP. Pourtant, dans nos prédictions cette position est systématiquement occupée par une asparagine. Plus de 7 Å séparent les carbones β du ligand et de la position 190 expliquant probablement l'absence de lysine. En revanche, nous observons l'introduction d'acides aminés basiques à des positions inattendues. La mutation Q367K ne nous intéresse pas dans la mesure où cette position est exposée au solvant et ne contribue pas à la spécificité de l'interaction AsnRS :AspAMP.

Par contre, les mutations E164K, Q187K, M223K, A335K (voir alignement figure 5.11) présentent un intérêt particulier puisqu'aucun acide aminé basique n'apparaît à ces positions dans les séquences naturelles d'AspRS. L'analyse structurale de ces différents mutants révèle alors des changements majeurs de conformations du ligand. Ainsi, les différentes orientations du ligand vont influencer la séquence du site actif de l'AsnRS, introduisant des charges positives dans la poche de manière à ce qu'elles interagissent avec le groupement carboxylate du ligand. Les figures 5.12, 5.13 et 5.14 présentent quatre orientations différentes du ligand conduisant chacune d'elles à l'introduction d'une lysine ou d'une arginine à différentes positions du site actif. Ces différentes mutations semblent toutes conduire à la formation de ponts salins entre le ligand et le résidu muté expliquant la meilleure affinité pour l'AspAMP que pour l'AsnAMP.

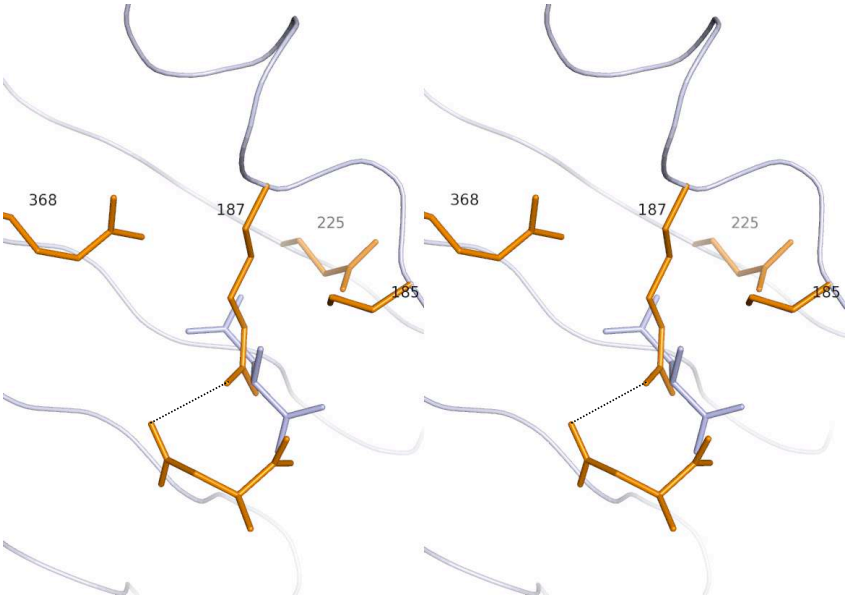


FIG. 5.12 – **Représentation stéréo de différents mutants présentant des orientations différentes du ligand.** Chacune de ces orientations conduit à l'introduction de charges positives à différents endroits de la protéine. La structure native de l'AsnRS de *T. thermophilus* est représentée en gris. Les ligands natif (gris) et mutant (orange) sont au centre du site actif. Chez le mutant (orange), le ligand interagit dans une orientation non native avec l'Arg 187. Les liaisons hydrogènes et les ponts salins sont matérialisés par des lignes pointillées.

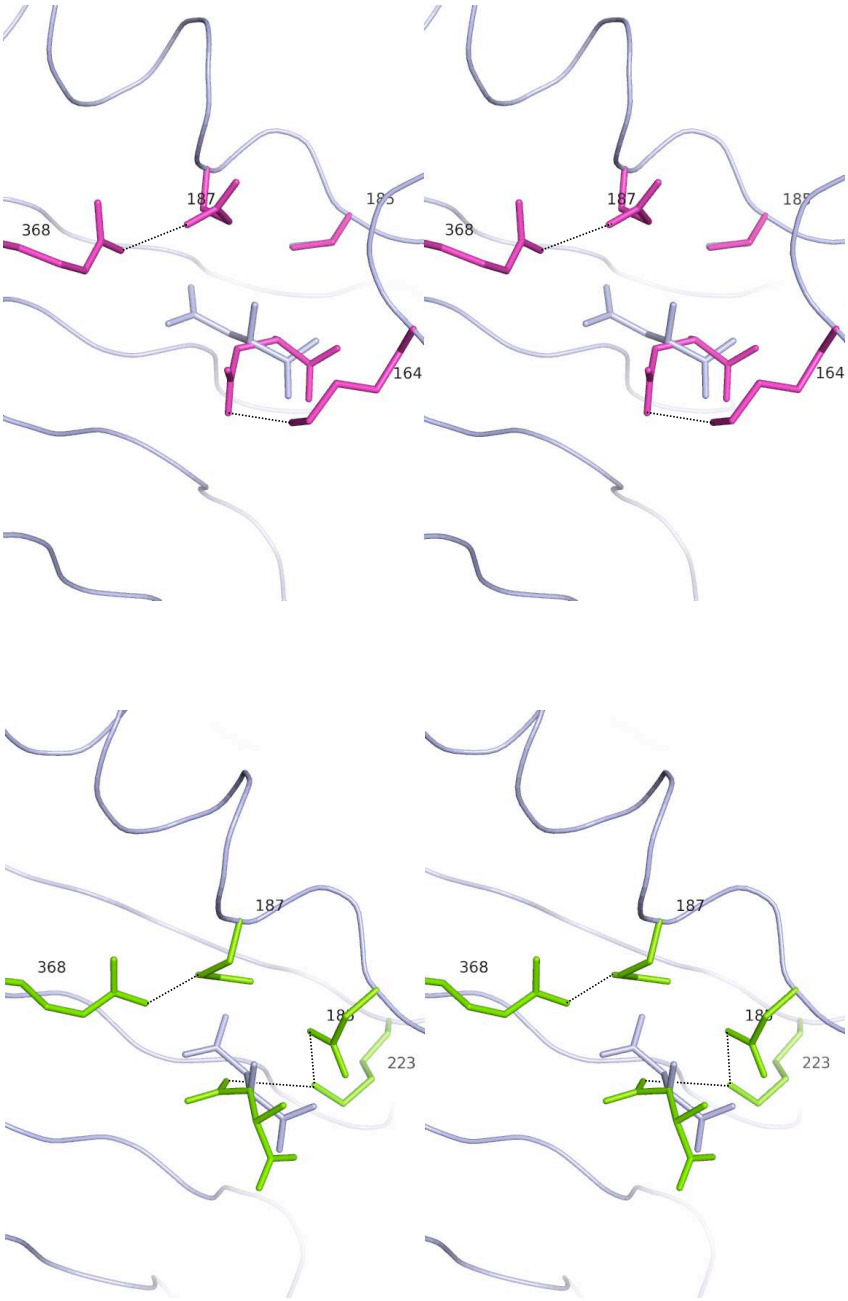


FIG. 5.13 – **Représentation stéréo de différents mutants présentant des orientations différentes du ligand.** Les conventions utilisées pour la figure 5.12 demeurent inchangées. En haut : dans la structure du mutant (magenta), le ligand se trouve dans une orientation complètement courbée et réalise un pont salin avec la Lys 164. Il est intéressant de remarquer que chez ces mutants, l'Arg 368 n'interagit plus avec la chaîne latérale du ligand mais crée une liaison hydrogène avec le résidu 187. En bas : chez le mutant (vert), le ligand interagit dans une orientation non native avec la Lys 223.

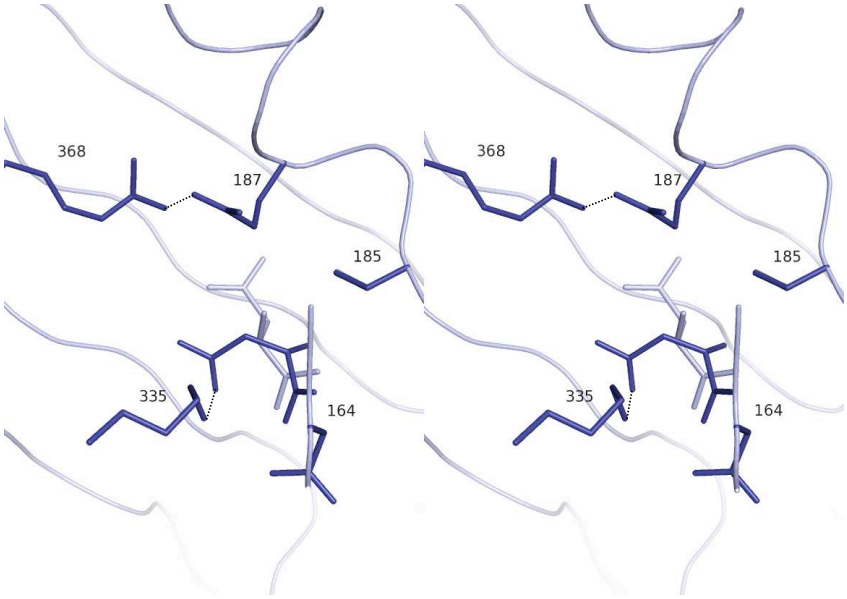


FIG. 5.14 – **Représentation stéréo de différents mutants présentant des orientations différentes du ligand.** Les conventions utilisées pour la figure 5.12 demeurent inchangées. Dans la structure du mutant (bleu), le ligand se trouve dans une orientation complètement courbée et réalise un pont salin avec la Lys 335.

Cependant, la réaction d'activation de l'acide aminé requiert l'alignement de celui-ci et de l'ATP [12], [28], [10]. L'acide aminé ne doit donc pas se présenter dans une orientation coudée par rapport à l'axe natif. Ceci n'est pas compatible avec la plupart des mutations obtenues, qui impliquent pour la majorité une orientation non native du ligand. En effet, sur les 86 mutants retenus, seuls 23 d'entre eux conservent l'orientation globale du ligand natif. Ces résultats ne remettent pas en cause pour autant la qualité de nos prédictions. En effet, si les mutations que nous prédisons sont visiblement incompatibles avec la réaction d'activation, elles peuvent cependant conférer à l'AsnRS une meilleure affinité pour l'AspAMP que pour l'AsnAMP. Les affinités respectives de chacun de ces 23 mutants pour l'AspAMP et l'AsnAMP sont présentées en table 5.1.

Par conséquent, nous avons recalculé une matrice d'énergie en contraignant l'orientation du ligand. . Par ailleurs, il fut très intéressant dans un premier temps de tester notre procédure de CPD à prédire les caractéristiques structurales et fonctionnelles des sites actifs de l'AspRS et de l'AsnRS en mutant une quarantaine de positions. Toutefois, notre but étant de réaliser de la mutagenèse il n'était pas raisonnable de proposer un trop grand nombre de mutations aux expérimentateurs.

Mutants	Affinité (kcal/mol)		Mutants	Affinité (kcal/mol)	
	AspAMP	AsnAMP		AspAMP	AsnAMP
mut72752	-23.61 ± 11.6	-15.22 ± 8.3	mut84329	-19.08 ± 8.2	-16.22 ± 8.4
mut112995	-22.29 ± 8.3	-17.14 ± 7.5	mut203074	-19.03 ± 9.3	-9.19 ± 8.1
mut54346	-22.11 ± 6.9	-21.74 ± 7.8	mut193628	-18.01 ± 7.9	-13.72 ± 8.9
mut6756	-21.33 ± 11.8	-12.78 ± 8.4	mut66644	-18.40 ± 7.4	-16.67 ± 9.5
mut34644	-20.98 ± 8.7	-7.93 ± 11.2	mut190753	-18.13 ± 8.2	-5.39 ± 7.8
mut60644	-20.87 ± 10.3	-18.57 ± 8.8	mut173966	-18.04 ± 6.8	-14.27 ± 12.8
mut92598	-20.71 ± 12.1	-19.40 ± 8.2	mut162370	-17.94 ± 9.5	-9.08 ± 8.4
mut174331	-20.62 ± 9.7	-11.04 ± 7.5	mut88144	-17.83 ± 7.3	-16.25 ± 7.9
mut194336	-20.44 ± 9.3	-14.37 ± 9.2	mut59264	-17.54 ± 7.9	-13.11 ± 6.7
mut65939	-20.34 ± 9.8	-19.88 ± 10.5	mut198373	-16.71 ± 8.6	-15.63 ± 9.2
mut131538	-19.80 ± 11.4	-6.39 ± 10.6	mut55010	-16.50 ± 7.3	-7.29 ± 8.7
mut185533	-19.18 ± 7.0	-17.58 ± 6.8			

Table 5.1 : Affinités de chacun des mutants pour l'AspAMP et l'AsnAMP en kcal/mol. Comme précédemment, nous observons un écart type important découlant des différents conformères caractérisant un mutant donné.

5.4 Design des 10 positions du site actif de l'AsnRS à proximité du ligand

5.4.1 Evolution dirigée

Les calculs ci dessus représentent un test intéressant de notre procédure, notamment pour sa capacité à prédire les caractéristiques structurales et fonctionnelles des sites actifs de l'AspRS et de l'AsnRS. Cependant, pour la suite, nous avons repris le problème avec un modèle légèrement plus simple. Le ligand n'est plus représenté par les 161 rotamères définis au départ mais par un jeu limité de respectivement 5 et 11 rotamères pour l'AspAMP et l'AsnAMP provenant de la bibliothèque de Tufféry [201]. Ces rotamères préservent l'orientation globale du ligand natif. De plus, nous avons fait le choix de nous limiter à un nombre de positions 'Actives' plus restreint. L'analyse des séquences et des structures des mutants obtenus dans nos études précédentes nous a permis d'identifier un jeu d'une dizaine positions à muter. Ces positions furent sélectionnées en partie en fonction de leur proximité avec le ligand ou de leur conservation au sein des AsnRS et AspRS. Avec ce modèle, nous avons généré 100000 séquences que nous avons classées selon leur score BLOSUM. Les 2000 meilleures furent reconstruites et classées en fonction de leur affinité pour l'AspAMP. Comme précédemment, l'affinité pour l'AspAMP des 100 meilleurs mutants fut de nouveau recalculée selon le protocole défini en section 5.3.

Ceci nous permit d'isoler 70 mutants ayant une meilleure affinité pour l'AspAMP que pour l'AsnAMP. La figure 5.15 présente les 10 positions mutées des mutants classés selon leur affinité pour l'AspAMP. Pour plus de clarté, seuls les 30 meilleurs sont représentés. Nous constatons que l'introduction d'acides aminés basiques se limite essentiellement aux positions 187 et 223. L'analyse structurale montre que la mutation M223K contribue principalement à stabiliser les oxygènes ot1 et o1p du ligand.

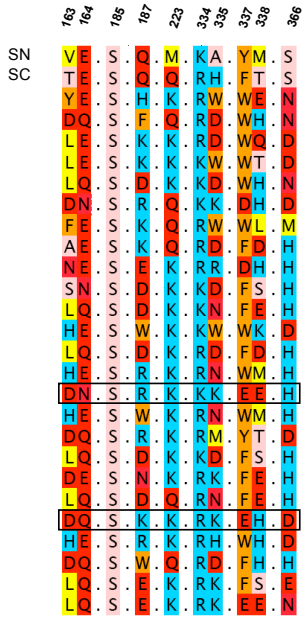


FIG. 5.15 – **Alignement de positions des séquences des mutants obtenus en présence de l'AspAMP dans le site actif.** La première ligne 'SN' représente la séquence native d'AsnRS de *T. thermophylus* tandis que la ligne suivante 'SC' correspond à la séquence consensus d'AspRS. Pour plus de clarté seuls 30 des 70 mutants sont représentés. Par ailleurs, seules les 10 positions 'Actives' sont présentes sur l'alignement. Globalement, nous introduisons un grand nombre de résidus chargés dans nos mutants. Les encadrés noirs présentent le cas extrême de mutants impliquant l'introduction de 5 nouveaux résidus chargés.

Par ailleurs, nos prédictions semblent préserver la stabilisation du groupement NH_3^+ du ligand. La Ser 185, interagissant avec le groupement ammonium du ligand dans la structure cristallographique, est parfaitement conservée au sein de nos séquences. De plus, dans la majorité des cas, cette stabilisation semble garantie par un pont salin impliquant le groupe NH_3^+ du ligand et un aspartate ou un glutamate en position 163 ou 164. Il est intéressant d'observer qu'on prédit généralement un aspartate en position 163 couplé à un résidu polaire en position 164 ou un glutamate en position 164 couplé à un résidu hydrophobe ou polaire en position 163. On note quelques rares cas ou aucune charge négative n'est retrouvée ni en position 163 ni en 164. Dans ces cas, on observe la présence d'un aspartate ou d'un glutamate en position 187 qui va restituer la stabilisation du groupement ammonium du ligand par l'intermédiaire d'un pont salin.

5.4.2 Analyse des mutants les plus prometteurs

Les 70 mutants furent analysés structuralement afin d'isoler une dizaine de candidats que nous testerons par dynamique moléculaire et ensuite expérimentalement. Il ne faut pas déstabiliser le site actif de l'AsnRS en introduisant trop de résidus chargés, ni s'éloigner de la charge nette de la poche. Par conséquent, en sélectionnant parmi les 70 mutant générés, nous avons voulu limiter le nombre de résidus chargés introduits dans le site actif à un résidu chargé positivement (l'aspartate étant chargé négativement) ou une paire de résidus de charges opposées. De plus, parmi les mutations introduisant de nouvelles charges, nous avons privilégié celles qui semblaient contribuer au changement de spécificité de l'enzyme. Par exemple, un grand nombre de mutants (parmi les 70) présentent une lysine ou une arginine en position 187 et 223. Dans ce cas, nous avons conservé uniquement la mutation impliquant la position 187. La mutation M223R/K semble participer à la stabilisation des oxygènes ot1 et o1p, rôle en parti déjà assuré par l'Arg 208 conservée chez toutes les AsnRS et maintenue 'Inactive' lors de nos calculs.

Ainsi, dans la plupart des cas, nous avons restauré la methionine en position 223. Dans le souci de réduire le nombre de mutations introduites, nous avons éliminé les mutations qui d'après l'analyse structurale, n'apportaient visiblement pas de gain de spécificité pour l'aspartate. Dans ces cas, l'acide aminé natif fut restauré. Ces étapes nous permirent d'isoler sept séquences présentant un site actif plus propice à recevoir l'aspartate que l'asparagine. La table 5.2 présente les séquences des sept mutants les plus prometteurs ainsi que leur énergies de liaison avec l'AspAMP et l'AsnAMP. Leurs structures sont représentées par les figures 5.16, 5.17, 5.18, 5.19, 5.20, 5.21 et 5.22.

	séquences										Affinité (kcal/mol)	
AsnRS	163	164	185	187	231	334	335	337	338	366	AspAMP	AsnAMP
AsnRS	V	E	S	Q	M	K	A	Y	M	S	-11.7	-19.3
AspRS	T	E	S	Q	Q	H	H	F	T	S	-	-
AspRS	169	171	193	195	223	448	449	451	452	487		
mut8	**F**	E	S	**R**	M	K	**W**	**W**	M	**D**	-25.4	-9.9
mut10479	**W**	E	S	**R**	M	K	A	**W**	M	**D**	-16.4	-5.1
mut7541	V	E	S	**K**	**Q**	K	A	**W**	**Q**	**D**	-20.2	-6.9
mut2299	V	E	S	**R**	M	K	**M**	Y	M	**D**	-27.9	-15.6
mut22992	V	E	S	**R**	M	K	A	Y	M	**N**	-22.2	-8.3
mut9085	**D**	**L**	S	**K**	**Q**	K	A	Y	M	S	-23.1	-5.2
mut9363	**D**	**V**	S	**N**	**K**	K	A	Y	M	S	-34.3	-18.3

Table 5.2 : Affinité moyenne des sept mutants pour l'AspAMP et l'AsnAMP. Les deux premières lignes indiquent respectivement la numérotation PDB des dix positions 'Actives' et la séquence native d'AsnRS de *T. thermophilus*. Les deux lignes suivantes correspondent à l'AspRS de *E. coli*. Les affinités présentées correspondent à des affinités moyennes calculées sur les 500 conformères de plus basse énergie de chaque mutant (pour plus de détail se référer à la section 5.3).

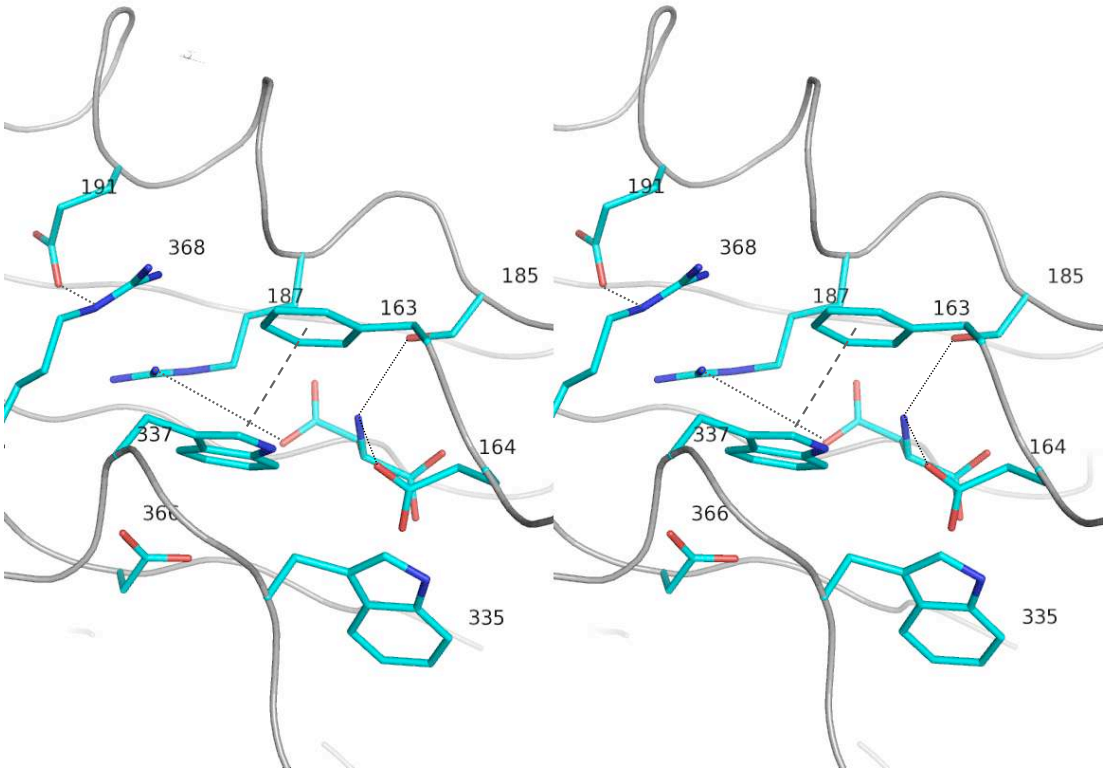


FIG. 5.16 – **Vue stéréo du site actif du mutant mut8.** Pour plus de clarté nous n'avons pas représenté la partie AMP du ligand commune à tous les mutants. On rappelle qu'elle a été maintenue fixé lors de nos calculs. Le squelette peptidique est représenté en gris en mode 'ribons'. Il correspond à celui de l'AsnRS de *T. thermophylus* et est commun à tous les mutants. Le ligand se trouve au centre du site actif. Les résidus d'intérêt sont représentés en 'sticks'. Les liaisons hydrogènes et ponts salins sont matérialisés par des lignes pointillées noires. Les interactions van der Waals sont représentées par une ligne plus épaisse. On note que l'orientation de l'Arg 368 a été modifiée puisqu'elle n'interagit plus avec le ligand. Dorénavant elle interagit avec le Glu 191. Ce phénomène sera observé chez tous les mutants présentés ci-dessous.

Le premier mutant (mut8) est représenté par la figure 5.16. Les acides aminés natifs 185 et 164 maintiennent la stabilisation du groupement ammonium du ligand. L'Arg 187 crée un pont salin avec le groupement carboxylate du ligand. Ainsi, la mutation Q187R pourrait contribuer à la spécificité de l'interaction AsnRS-aspartate. Par ailleurs, l'Arg 187 semble interagir avec l'Asp 366 ce qui pourrait la stabiliser en absence du ligand. D'autre part ce mutant qui présente un total de cinq

mutations comble une partie de l'espace inoccupé de la poche par la présence de cycles aromatiques. Nous pouvons espérer que ces derniers contribueront à augmenter la stabilité de notre mutant via des interactions hydrophobes favorables et des interactions van der Waals entre la Phe 163 et le Trp 337.

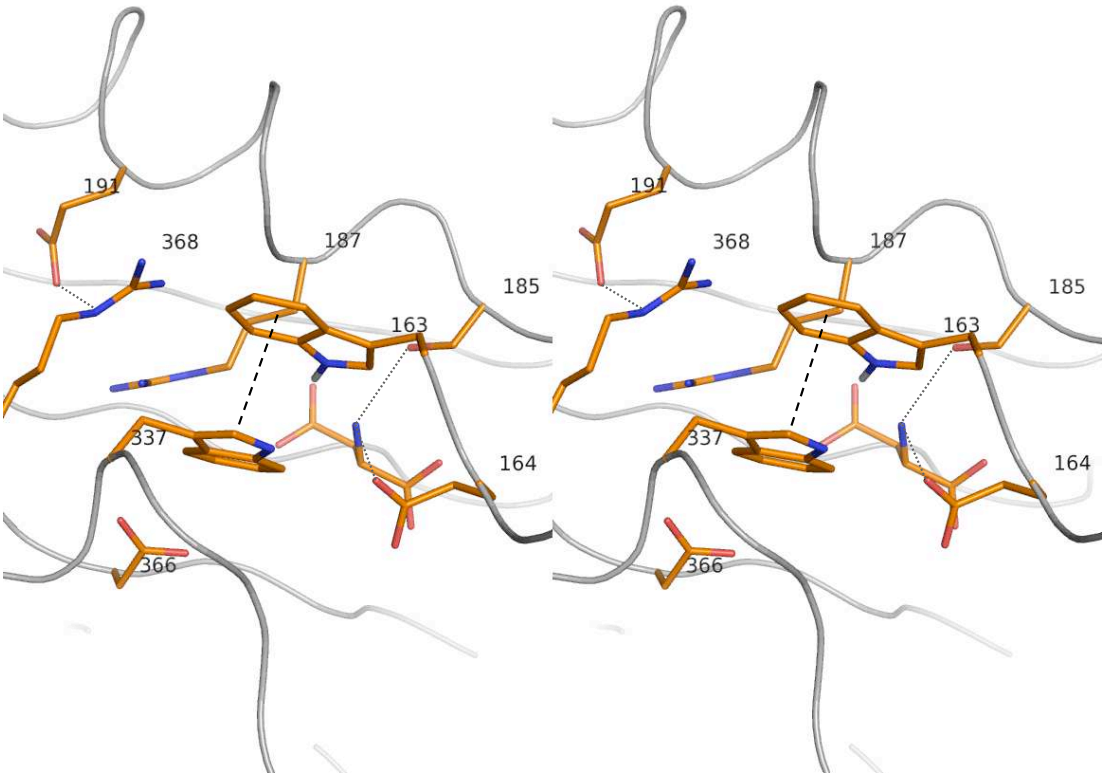


FIG. 5.17 – **Vue stéréo du site actif du mutant mut10479.** Le code de représentation est identique à celui de la figure 5.16.

Le second mutant (mut10479) présente de fortes similitudes avec le premier (figure 5.17). La spécificité de l'interaction semble assurée par le pont salin entre l'Arg 187 et le groupe carboxylate du ligand. Cependant, la position 335 conserve son alanine native contrairement au mutant mut8 qui contenait la mutation A335W. En comparant ces deux mutants, nous voulons évaluer l'influence du tryptophane en position 335 sur la stabilité de la protéine.

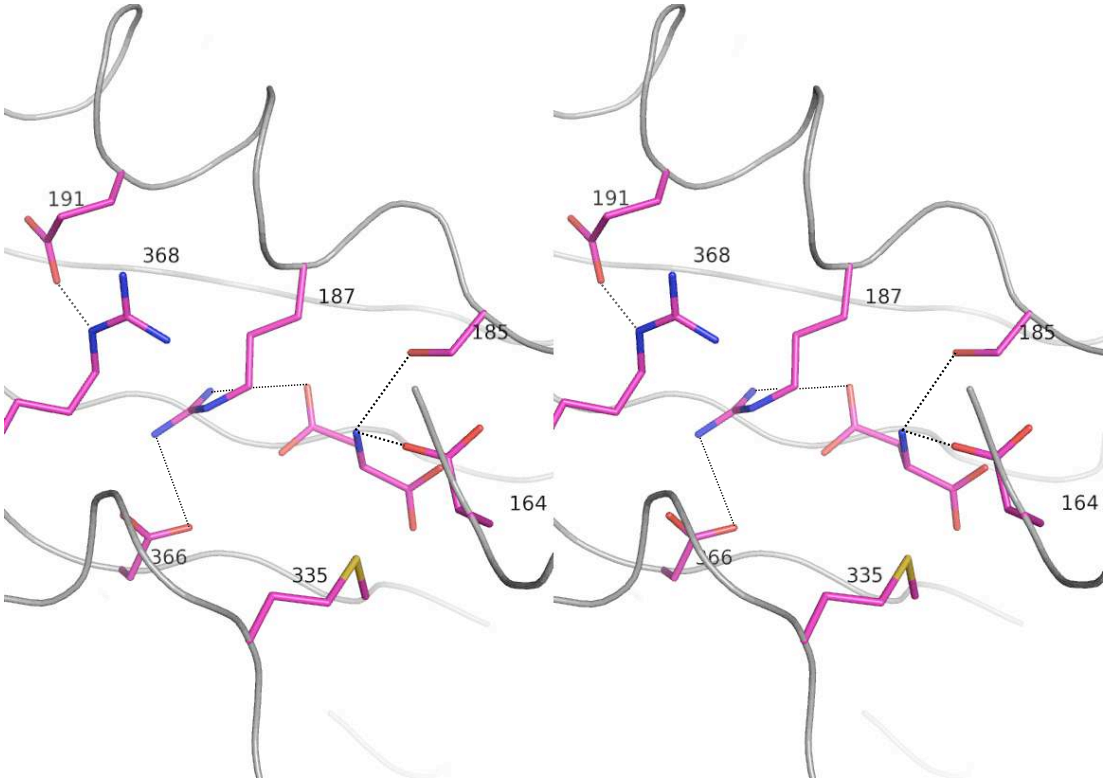


FIG. 5.18 – **Vue stéréo du site actif du mutant mut2299.** Le code de représentation est identique à celui de la figure 5.16.

Avec le mutant suivant (mut2299), nous testons toujours l'influence des résidus aromatiques dans le site actif sur la stabilité de la protéine (figure 5.18). Ce mutant à l'inverse des deux précédents n'introduit aucun nouveau résidu aromatique. Il introduit cependant une méthionine (résidu hydrophobe) en position 335. La spécificité est encore garantie par l'Arg 187, elle même stabilisée par l'Asp 366. Par ailleurs, rappelons que le groupement ammonium est toujours stabilisé par les résidus natifs 164 et 185.

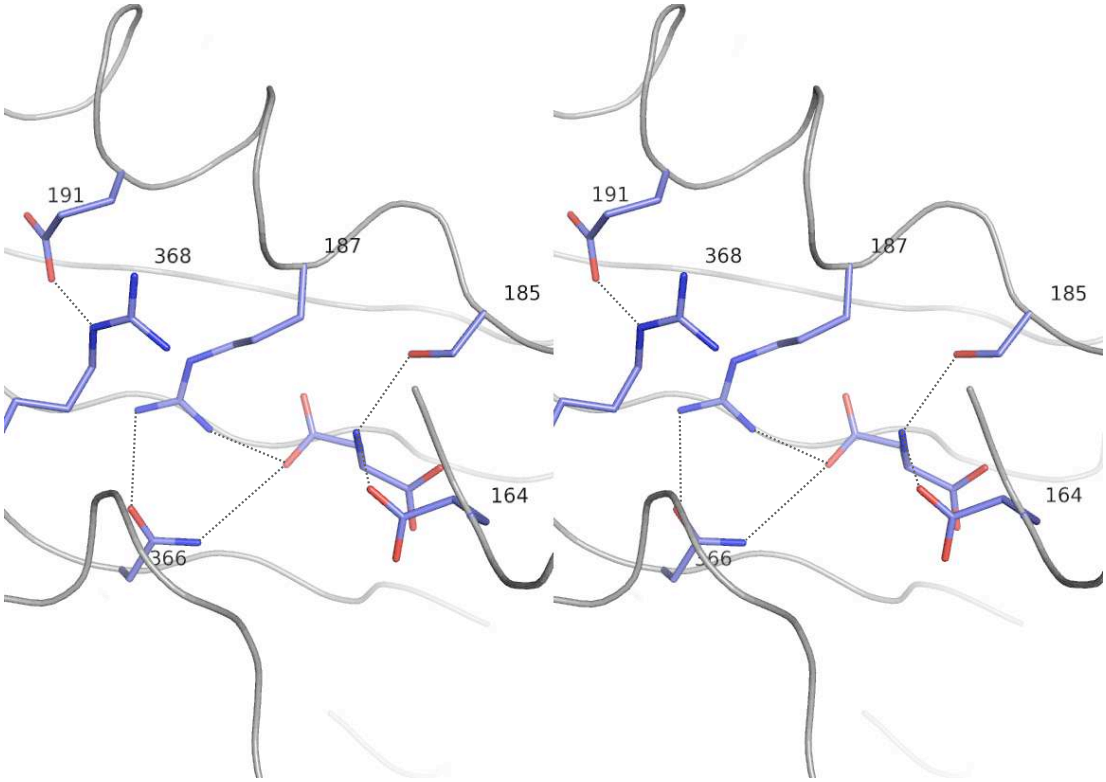


FIG. 5.19 – **Vue stéréo du site actif du mutant mut22992.** Le code de représentation est identique à celui de la figure 5.16. La chaîne latérale du ligand interagit avec le groupement guanidium de l'Arg 187 et réalise une liaison hydrogène avec l'Asp 366.

Les mutants suivants excepté le mut7541, présentent tous l'introduction d'une seule charge positive supplémentaire (les trois mutants précédents présentaient deux charges supplémentaires de signes opposés conduisant à une conservation de la charge globale du site actif).

Le double mutant (mut22992) a l'intérêt de présenter une double spécificité pour l'aspartate (figure 5.19). En effet le groupement carboxylate du ligand semble interagir simultanément avec l'Arg 187 et l'Asn 366. Son groupement NH_3^+ est maintenu comme pour les autres mutants par des interactions électrostatiques avec les résidus natifs 164 et 185. Enfin, ce mutant présente l'avantage de n'impliquer que deux mutations ponctuelles.

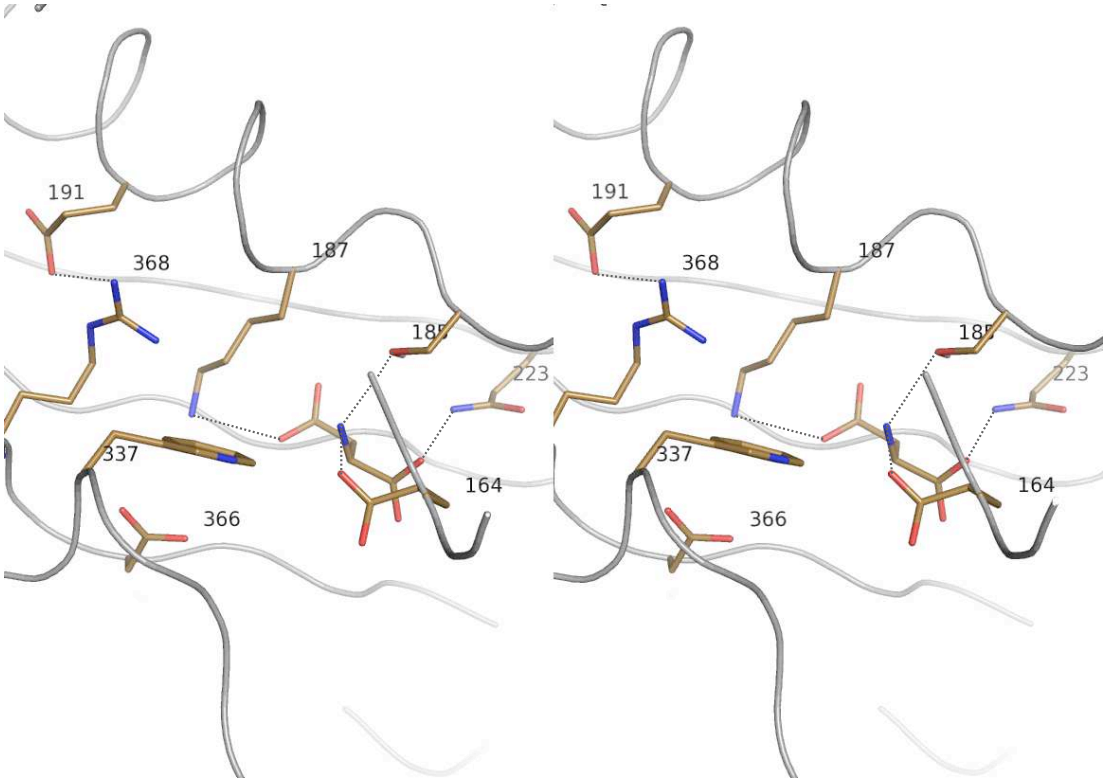


FIG. 5.20 – **Vue stéréo du site actif du mutant mut7541.** Le code de représentation est identique à celui de la figure 5.16. L'Arg 187 prédite chez tous les mutants précédents est remplacée par une lysine.

Avec le mutant mut7541, nous comparons l'influence d'une lysine en position 187 à la place de l'arginine prédite à cette même position chez les mutants précédents (figure 5.20). Tout comme l'arginine, la lysine interagit avec le groupement carboxylate du ligand et l'Asp 366. Ce mutant présente une nouvelle mutation en position 223 avec l'introduction d'un résidu polaire participant à la stabilisation de l'ot1 du ligand.

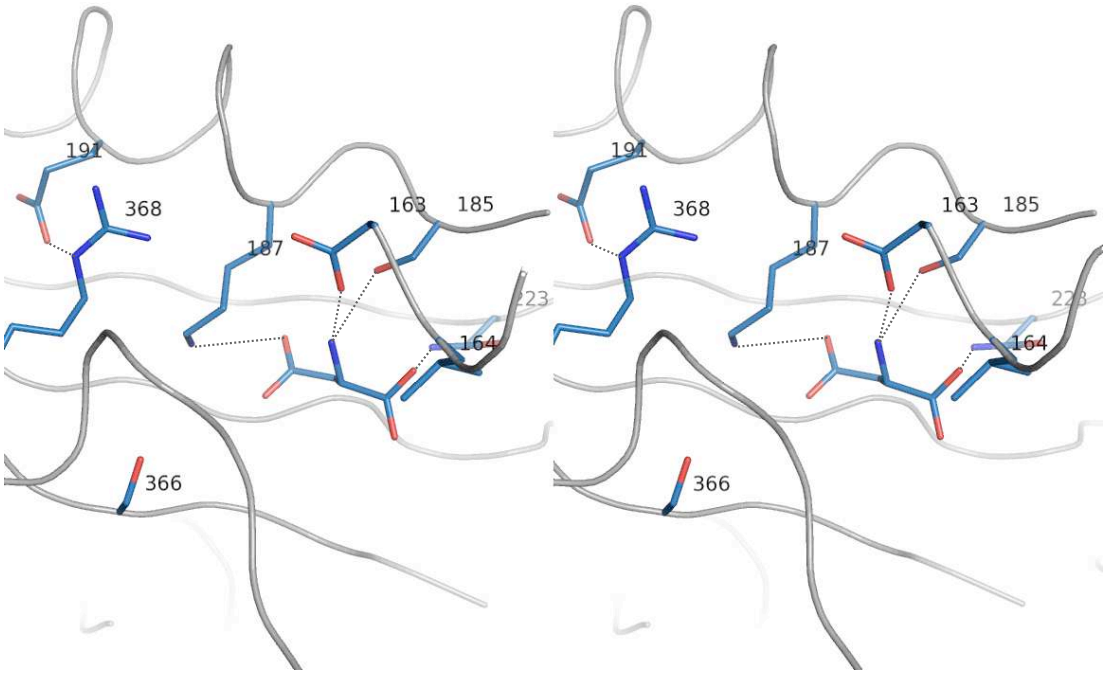


FIG. 5.21 – **Vue stéréo du site actif du mutant mut9085.** Le code de représentation est identique à celui de la figure 5.16. Il est intéressant de comparer l'orientation de la Lys 187 chez ce mutant et chez mut7541. En absence d'Asp 366, la Lys 187 s'oriente vers le centre du site actif pour interagir exclusivement avec le ligand.

Le mutant mut9085 présente aussi une lysine en position 187 (figure 5.21). Il est intéressant de remarquer l'orientation de la lysine tournée vers l'intérieur du site en absence de la mutation S366D. En effet, en présence d'aspartate en position 366 la lysine s'oriente de manière à interagir simultanément avec l'Asp 366 et le ligand (voir mutant mut7541). Dans le cas présent, la lysine interagit uniquement avec le groupement carboxylate de l'aspartate. Il sera alors intéressant de comparer les affinités respectives des mutants mut7541 et mut9085. Par ailleurs, la stabilisation du groupement ammonium est assurée par l'Asp 163, la position 164 étant occupée par une leucine.

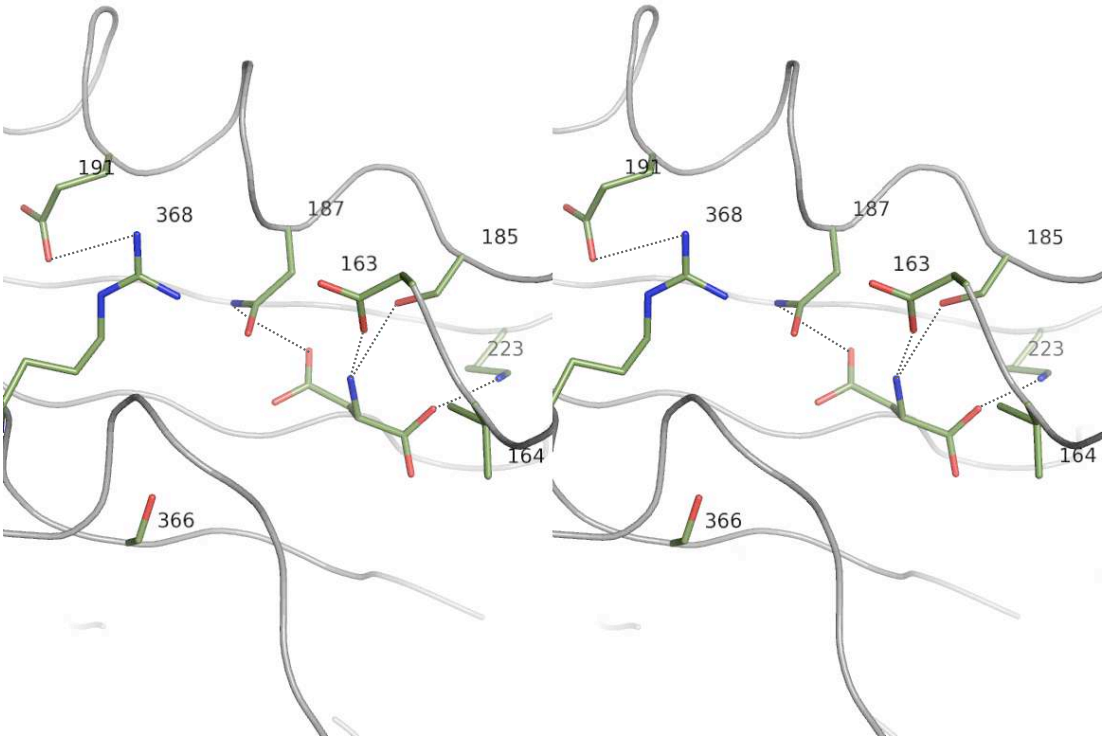


FIG. 5.22 – **Vue stéréo du site actif du mutant mut9363.** Le code de représentation est identique à celui de la figure 5.16. Chez mut9363, la position 187 est occupée par un acide aminé polaire à la différence des autres mutants proposant tous un acide aminé chargé positivement.

Enfin, le mutant mut9363, présente une asparagine en position 187 (figure 5.22). Cette dernière pourrait alors stabiliser le ligand par l'intermédiaire d'interactions hydrogènes entre son groupement NH_2 et le groupe carboxylate de l'aspartate.

La stabilité de ces mutants ainsi que leurs affinités respectives pour l'AspAMP et l'AsnAMP sont actuellement évaluées par des simulations de dynamique moléculaire. Les mutants les plus prometteurs seront ensuite proposés aux expérimentateurs. Cette étude présentait une certaine difficulté dans la mesure où notre objectif n'était pas uniquement de modifier l'affinité de l'aaRS pour son ligand non natif mais aussi de modifier l'activité catalytique de l'enzyme de façon à ce que l'aaRS

soit capable d'aminoacyler en quantité raisonnable des ARNt avec son ligand non natif. Seulement, ceci implique que l'activité de l'enzyme ne soit pas détériorée par nos mutations et que la réaction d'activation ait lieu. De plus ce système est complexe puisqu'il interagit simultanément avec plusieurs substrats (acide aminé, ATP, ARNt), catalyse la réaction d'aminoacylation en plusieurs étapes... Par ailleurs, bien que les sites actifs de l'AsnRS et l'AspRS présentent de grandes similitudes, chaque enzyme est très spécifique de son acide aminé respectif. Tous ces faits rendent difficile l'ingénierie de ces deux synthétases. Néanmoins, nous avons montré que notre procédure était sensible à la nature du ligand dans la poche et était capable de restituer des caractéristiques structurales mais aussi fonctionnelles des AspRS et AsnRS.

Chapitre 6

Conclusion

Depuis quelques années, la conception des protéines est un domaine en plein essor. Elle présente de nombreuses applications telles que le développement de protéines thermostables, l'ingénierie de nouveaux ligands, biosenseurs et catalyseurs et joue un rôle de plus en plus important dans les domaines pharmaceutiques et bio-technologiques. Les approches expérimentales sont généralement limitées par le temps et le coût des expériences. L'approche computationnelle répond à ce problème en parcourant de façon exhaustive et contrôlée l'espace des séquences et des conformations et permet par ailleurs d'explorer des possibilités jusque là inenvisagées. Nous nous sommes donc intéressés au problème du CPD.

Au cours de cette étude, nous avons mis en place une procédure automatique de CPD reposant sur une fonction d'énergie relativement simple. Cette fonction d'énergie utilise une représentation du solvant (modèle CASA) et de l'état déplié (modèle des tripeptides) très simplifiée. Néanmoins, elle s'avère robuste puisqu'elle a fait ses preuves, sous réserves de légers ajustements, dans des applications très diverses [134], [185], [186]. En particulier, nous avons montré sa capacité à prédire des changements de stabilité ou d'affinité protéine-ligand associés à des mutations ponctuelles. Les résultats obtenus sont de qualités comparables à celles obtenues avec des modèles de solvant plus complexes tels que le modèle GB/ACE [134], [186].

Par ailleurs, nous avons évalué notre procédure de CPD pour le *design* de différentes protéines globulaires [185], [186]. Les séquences prédites présentent des scores d'identité avec la séquence native de 35% en moyenne, score nettement supérieur à celui obtenu par le groupe de Wodak (23.9%) [98]. Ceux obtenus par Pokala et Handel s'échelonnent entre 33,5 et 46,7% [165]. Néanmoins, leur fonction d'énergie utilise le modèle de GB bien plus gourmand que celui de CASA. Les résultats obtenus par le groupe de Baker sont comparables aux notres (37%) [180]. Cependant, il faut souligner que leur fonction d'énergie fut optimisée pour reproduire des séquences de type natif, ce qui n'était pas notre cas.

Ensuite, nous nous sommes intéressés à l'évolution dirigée de l'AsnRS, l'objectif étant de redessiner son site actif de manière à ce qu'elle lie préférentiellement l'aspartate au détriment de son acide aminé natif, l'asparagine. Nos études précédentes montrèrent que d'une application à l'autre et d'un système à l'autre, la fonction d'énergie nécessitait des ajustements même si la procédure générale était conservée. Le site actif de l'AsnRS s'avère plus complexe puisqu'il est régi par des contraintes structurales devant aussi répondre à des nécessités fonctionnelles. Ainsi, l'AspRS, très proche de l'AsnRS, fut utilisée pour ajuster et évaluer la fonction d'énergie dans sa capacité à reproduire des séquences à caractère natif ou à prédire des changements d'affinité associés à des mutations ponctuelles. Les séquences de plus basse énergie produites par notre algorithme présentent des scores d'identité avec la séquence native d'environ 25%. En les classant selon leurs score BLOSUM, les meilleures présentent jusqu'à 40% d'identité avec la séquence native. Ce résultat est satisfaisant dans la mesure où nous nous sommes concentrés sur un site actif, région de la protéine difficile à modéliser. Dans ces régions, les résidus obéissent à des lois physiques afin de garantir une stabilité minimale à la protéine, mais sont aussi régis par la nécessité de conférer une fonction précise à la protéine (reconnaissance de substrat, catalyse). Or notre fonction d'énergie repose principalement sur des potentiels physiques. Toutefois, nous avons montré que nous étions capables de prédire des séquences présentant des caractéristiques structurales mais aussi fonctionnelles

de la protéine. En particulier, nous rappelons le cas de la Lys 198 (numérotation AspRS d'*E. coli*), prédite par notre algorithme lorsque les calculs furent réalisés en présence d'AspAMP dans la poche et absente de nos prédictions lors des calculs effectués avec l'AsnAMP. Cette lysine conservée au sein de toutes les AspRS, joue un rôle déterminant dans la spécificité de l'interaction Asp :AspRS. Ainsi, en introduisant une contrainte biologique (nature du ligand), nous influençons le profil des séquences obtenues, particulièrement celui des positions en interaction directe avec le ligand. Ce résultat était très encourageant pour l'ingénierie d'une synthétase de spécificité modifiée.

Dans le chapitre 5, nous avons réalisé l'évolution dirigée du site actif de l'AsnRS. Notre procédure de CPD permit d'isoler une dizaine de mutants qui d'après nos calculs présentent une meilleure affinité pour l'AspAMP que pour l'AsnAMP. L'analyse structurale de ces derniers montre que notre fonction d'énergie a optimisé la stabilité de la protéine en favorisant l'enfouissement de résidus hydrophobes tout en prédisant des mutations jouant un rôle fonctionnel dans l'enzyme. Ces dernières semblent favoriser la liaison de l'AspAMP au détriment de l'AsnAMP bien que ces prédictions restent à prouver par des mesures expérimentales.

Dans un futur proche, il sera nécessaire d'introduire dans notre algorithme d'optimisation, une compétition entre le ligand d'intérêt et d'autre(s) ligand(s) tout en garantissant une stabilité minimale. Cela permettrait de rendre compte de la spécificité de l'interaction du complexe. Ainsi, on produirait des mutants optimisés sur leur pouvoir discriminatoire plutôt que sur leur stabilité. Par ailleurs, cela éviterait les nombreux filtres que nous avons du mettre en place. Bien que le modèle de solvant CASA ait fait ses preuves dans de nombreuses applications, il serait aussi intéressant d'implémenter le modèle GB dans notre fonction d'énergie. Nos études précédentes ont montré que pour certaines applications, le modèle GB/HCT fournissait des résultats de meilleure qualité [134], [186]. Enfin, dans le cas des interactions protéine-ligand ou protéine-protéine, la protéine peut subir des changements conformationnels incluant des réajustements de son squelette peptidique.

Ainsi, pour mieux aborder le *design* de complexes, il conviendrait d'incorporer de la flexibilité dans le squelette peptidique. Néanmoins, nous avons montré qu'avec un modèle très simple, nous avons obtenu des résultats très encourageants dans le domaine ambitieux de l'évolution dirigée.

Chapitre 7

Annexe

Computational Protein Design: Software Implementation, Parameter Optimization, and Performance of a Simple Model

MARCEL SCHMIDT AM BUSCH, ANNE LOPES, DAVID MIGNON, THOMAS SIMONSON
Laboratoire de Biochimie (CNRS UMR7654), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France

Received 13 July 2007; Revised 28 September 2007; Accepted 7 October 2007
DOI 10.1002/jcc.20870
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: Computational protein design will continue to improve as new implementations and parameterizations are explored. An automated protein design procedure is implemented and applied to the full redesign of 16 globular proteins. We combine established but simple ingredients: a molecular mechanics description of the protein where nonpolar hydrogens are implicit, a simple solvent model, a folded state where the backbone is fixed, and a tripeptide model of the unfolded state. Sequences are selected to optimize the folding free energy, using a simple heuristic algorithm to explore sequence and conformational space. We show that a balanced parametrization, obtained here and in our previous work, makes this procedure effective, despite the simplicity of the ingredients. Calculations were done using our Proteins @ Home distributed computing platform, with the help of several thousand volunteers. We describe the software implementation, the optimization of selected terms in the energy function, and the performance of the method. We allowed all amino acids to mutate except glycines, prolines, and cysteines. For 15 of the 16 test proteins, the scores of the computed sequences were comparable to those of natural homologues. Using the low energy computed sequences in a BLAST search of the SWISSPROT database, we could retrieve natural sequences for all protein families considered, with no high-ranking false-positives. The good stability of the designed sequences was supported by molecular dynamics simulations of selected sequences, which gave structures close to the experimental native structure.

 J Comput Chem 00: 000–000, 2007

Key words: protein engineering; molecular mechanics; inverse protein folding; distributed computing; combinatorial optimization

Introduction

With the development of genomics and rapidly growing databases of protein sequences, protein structure prediction is increasingly important. The protein folding problem, or prediction of structure from sequence alone, remains a major challenge. In the 80's, the inverse folding problem was formulated: instead of predicting the 3D structure from the sequence, one considers a given backbone structure and predicts the amino acid sequences that fold into it.[1–4] Computational protein design addresses this inverse problem. The methods developed over the last two decades and their applications[4–26] represent rigorous tests of our understanding of the mechanisms that shape protein sequences and structures. They provide not only tools for engineering new proteins,[27–32] but also methods for structure prediction. Indeed, protein design can be seen as an evolutionary model that evaluates the sequence evolution within a given protein family when a structural or functional constraint is applied.[15,33–35] A 25–30% sequence homology between proteins is generally sufficient for them to share a common fold, and computational design methods have the potential to become a fold-recognition method that can categorize unknown proteins into previously determined fold families.[15,22,34,35] As new implementations and parameterizations are explored, computational protein design will continue to improve.

We use simple, existing ingredients to implement a method whose performance appears to be competitive with several existing, more sophisticated methods.[10,36] We benefit from our previous testing and optimization of a simple implicit solvent model, and we describe the further optimization of other parameters in the

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages0192-8651/suppmat

Correspondence to: T. Simonson; e-mail: thomas.simonson@polytechnique.fr

Contract/grant sponsor: Agence Nationale pour la Recherche

energy function. Folded conformations have the experimental, native backbone conformation, in contrast to some recent methods that use multiple backbone conformations.[8,11,15,22–24] Sidechains occupy rotamers from a simple, backbone-independent rotamer library.[37] Sequences are selected to optimize the folding free energy, using a simple, heuristic algorithm to explore sequence and conformational space.[10] The method employs a molecular mechanics description of the protein, with a force field that treats nonpolar hydrogens implicitly.[38] Solvent is described implicitly, using a very simple model that includes a screened Coulomb energy and a solvent-accessible surface energy. This Coulomb/Accessible Surface Area, or CASA model was recently parameterized and tested for protein stability and for sidechain reconstruction, which are key steps in protein design.[39] The unfolded state model is also very simple, with each amino acid interacting only with nearby backbone groups (tripeptide model) and with solvent. An additional, empirical correction is derived here (similar to earlier work[9,11,40]), which depends on the amino acid type, and is chosen to provide a realistic overall amino acid composition. Energy calculations are performed using our Proteins @ Home distributed computing platform (biology.polytechnique.fr/proteinsathome), which will be described in detail elsewhere.[41]

We test the capabilities of the model for the nearly-complete redesign of 16 proteins (only glycines, prolines, and cysteines are not allowed to mutate). The test set includes eight SH3 domains and a diverse set of eight other proteins, from eight different families in the SCOP classification.[42] A longer-term goal is to study the inverse folding problem and the use of protein design for protein fold recognition.[22,41] For this, we want to produce sets of sequences that are realistic but also large and diverse. This contrasts with some applications where only a few, highly-optimized sequences are sought. For 15 of the test proteins, the computed sequences have similarity scores comparable to natural homologues. They also display a good structural stability when subjected to molecular dynamics simulations using a high quality, generalized Born, implicit solvent model.[39,43]

Overall, the method implemented here is similar to several existing methods but uses somewhat simpler ingredients. For example, our method ressembles that proposed by Wodak and coworkers,[10,36] but uses a simpler force field and a simpler rotamer library. Nevertheless, our implementation achieves a performance that is distinctly improved compared to Wodak et al. and appears to be competitive with several other, still more complex implementations, including methods that use a flexible backbone and/or a generalized Born solvent. The improvements can be partly attributed to our reparametrization of the CASA solvent model[39,44] and a simple but carefully optimized unfolded state model.

Methods

Folded and Unfolded States

Sequences and structures are selected based on their folding free energies, ΔG_{fold}, the difference between the free energy of their folded and unfolded states. In the folded state, the coordinates of the protein backbone are kept fixed, while sidechains occupy rotamers from the backbone-independent Tuffery library.[37] The backbone conformation was obtained by subjecting the protein crystal structure to 500 steps of conjugate gradient energy minimization. During the minimization, the effect of solvent was represented by a uniform dielectric constant of 20, applied to the Coulomb electrostatic energy term. The minimization led to an rms deviation from the experimental structure of 0.56—0.90 Å (depending on the protein) and a protein radius of gyration about 0.1 Å smaller than the crystal structure.

In the unfolded state, the amino acid sidechains do not interact with each other, but only with nearby backbone and with solvent (through the CASA implicit solvent model). Specifically, for each amino acid type X, we considered a large number of possible tripeptide structures with the sequence Ala-X-Ala, with backbone geometries taken from five proteins (lysozyme, ribonuclease A, bovine pancreatic trypsin inhibitor, Staphylococcal nuclease, and the α-toxin). The lowest-energy combination of backbone structure and sidechain rotamer was taken to represent the preferred structure of X in the unfolded state. The corresponding energy, E_{X}, represents the contribution of X to the unfolded state free energy. An additional (and smaller) contribution, e_{X}, was determined empirically, so as to obtain reasonable overall amino acid compositions in the final computed sequences; the optimization of e_{X} is described later (Empirical correction to the unfolded state energy section). For a given amino acid sequence, the unfolded state free energy is obtained by summing the contributions $E_{\text{X}} + e_{\text{X}}$ of the individual amino acids.

In protein design, we perform rounds of random mutagenesis, transforming a given sequence A into a new sequence B. By comparing the folding energies for A and B, we can determine which sequence is most favorable. Because our energy function is pairwise additive (see later), and because the backbone structure is fixed in the folded state, we account correctly for ΔG_{fold} if we include all pairwise interactions between sidechains and between each sidechain and the backbone. In particular, interactions between different portions of the protein backbone cancel when two sequences are compared, both in the folded or the unfolded state, so that no important interactions are missed through the tripeptide unfolded model.

Effective Energy Function

The effective energy function was described in detail elsewhere.[39] Briefly, we use the Charmm19 molecular mechanics energy function[38] along with the CASA implicit solvent model. With CASA, the solvent contribution is the sum of a screened Coulomb term and a solvent accessible surface term:

$$E_{\text{solv}} = \left(\frac{1}{\epsilon} - 1\right) E_{\text{coul}} + \alpha \sum_i \sigma_i A_i. \qquad (1)$$

Here, E_{coul} is the usual Coulomb energy, ϵ is a dielectric constant, the righthand sum is over the protein atoms i, A_i is the solvent accessible surface area of atom i, σ_i is an atomic solvation coefficient (measured in kcal/mol/Å^2), which depends on the atom type, and α is an overall scaling factor for the surface term.

Interactions between distant groups were omitted through the following cutoff scheme. If the inter-C$_\beta$ distance was above 15 Å

(respectively, below 10 Å), a residue pair was omitted (included). Otherwise (inter-C_β distance between 10 and 15 Å), if the minimum inter-sidechain distance was 9 Å or less, the pair was included.

Surface areas were computed using the Lee and Richards algorithm,[45] implemented in the XPLOR program,[46] using a 1.4 Å probe radius. For reasons of efficiency, following Street and Mayo,[47] we assume that A_i can be obtained by summing the contact areas A_{ij} between atom i and its neighbors j, and subtracting the contact, or solvent-inaccessible area $C_i = \sum_j A_{ij}$ from the total area of atom i. This approximation has the enormous advantage that the surface energy takes the form of a sum over pairs of amino acids. However, it leads to a systematic error, since the contact areas can overlap: a portion of atom i can be in contact with two atoms j and j' at a time. Street and Mayo showed, and we confirmed[39] that the systematic error can be largely corrected by applying a scaling factor of 0.5 to contact areas A_{ij} that involve at least one buried atom (i or j); for details, see.[39]

The interaction energy between each pair of sidechains, or between a sidechain and the backbone, involved a short energy minimization stage.[10] Each sidechain was first subjected to 15 steps of Powell minimization, with the backbone fixed and inter-sidechain interactions excluded. Then, interactions between the sidechain pair were included and a further 15 steps of minimization performed. The sidechain interaction energy was taken from this last, minimized structure.

The atomic solvation coefficients σ_i are the ones used in our previous work: 0.012 kcal/mol/Å^2 for carbons and sulfur; -0.06 kcal/mol/Å^2 for oxygen and nitrogen; zero for hydrogens, and -0.15 kcal/mol/Å^2 for ionized groups.[39] In the previous work, we did extensive testing and comparison of several different sets of surface parameters, based on sidechain reconstruction, protein solvation energies, and mutations of over 1000 sidechains (including buried sidechains).[39] Thus, the atomic solvation coefficients and the surface calculations used here can be viewed as extensively optimized and tested.

Sequence Optimization

We used a heuristic optimization procedure developed by Wernisch et al.[10] One of the goals of this work is to continue to test the performance of this method. A "heuristic cycle" proceeds as follows: an initial amino acid sequence and set of sidechain rotamers are chosen randomly. These are improved in a stepwise way. At a given amino acid position i, the best amino acid type and rotamer are selected, with the rest of the sequence held fixed. The same is done for the following position $i+1$, and so on, performing multiple passes over the amino acid sequence until the energy no longer improves (or a set, large number of passes is reached). The final sequence, rotamer set, and energy are output, ending the cycle. The method can be viewed as a steepest descent minimization, starting from a random sequence, and leading to a nearby, local, (folding) energy minimum. For the design calculations below, we typically perform ~450,000 heuristic cycles for each protein, thus sampling a large number of local minima on the energy surface. Cysteines, glycines, and prolines are expected to have a special effect on the protein's folded and unfolded state structures, which may not be accurately captured by our method. Therefore, if these amino acids are present in the native sequence, they are not mutated; all other amino acids are allowed to mutate freely (but not into Cys, Gly, or Pro).

Empirical Correction to the Unfolded State Energy

Given the simplicity of the effective energy and the unfolded state model, it was necessary to add an empirical correction to the unfolded state energies. For each amino acid I, a correction e_X is defined, where X represents the current amino acid type at position I. There are 17 values to optimize, corresponding to the 17 amino acid types that are free to mutate. We optimized the e_X so as to obtain reasonable overall amino acid compositions for the designed sequences of four test proteins: Grb2 (1gcqB), Vav (1gcqC), c-Src (1cka), and SHG (1shg), which are all SH3 domains (see Table 1). Homologous proteins were found by a BLAST search of the SWISSPROT database, using an identity threshold of at

Table 1. Test Set of Proteins.

PDB	Short name	Full name	SCOP family
1gcqB	Grb2	Growth factor receptor-Bound protein 2	SH3 domain
1gcqC	Vav	Vav proto-oncogene	SH3 domain
1cka	c-Crk		SH3 domain
1shg	Alpha-spectrin		SH3 domain
1abo	ABL kinase	ABL tyrosine kinase	SH3 domain
1ad5	HCK kinase	Haematopoietic cell kinase	SH3 domain
1csk	Csk	c-Src specific tyrosine kinase	SH3 domain
1fmk	c-Src	c-Src tyrosine kinase	SH3 domain
1lz1	Lysozyme		C-type lysozyme
4pti	BPTI	Pancreatic trypsin inhibitor	Kunitz-type inhibitors and BPTI-like toxins
1ctf	L7/L12	Ribosomal protein L7/L12	ribosomal protein L7/L12, C-terminal domain
1enh	Homeodomain	Engrailed homeodomain	homeodomain
1pgb	Protein G	Protein G, Ig binding domain	immunoglobulin binding domains
1zaa	Zif268		classic zinc finger, C2H2
1c9o	CSP	Cold shock protein	cold shock DNA-binding domain-like
1bdd	Protein A	Protein A, B-domain	Ig binding protein A modules

least 35% to the native, query protein, for a chain length no less than 90% of the native length. The mean amino acid frequencies $f_X^{\exp}$ were computed by averaging over this data set. We then proceeded iteratively, with the e_X initially set to zero. At each iteration, 30,000 sequences were computed for each protein (through 30,000 heuristic cycles). The corresponding amino acid frequencies, f_X^{calc}, averaged over all sequences, proteins, and amino acid positions, were compared to the experimental frequencies $f_X^{\exp}$. The energy correction e_X was then modified according to the Boltzmann-like relation:

$$e_X^{\text{new}} = e_X^{\text{old}} + 0.5 \ln \frac{f_X^{\exp}}{f_X^{\text{calc}}}. \quad (2)$$

With this scheme, if a given type X is too abundant in the designed sequences, eq. (2) leads to an increased stability of the unfolded state when X is present, so that X will be less abundant in the next round. After eight rounds, the frequencies converged to the experimental values; the corrections e_X no longer changed significantly, and the procedure was stopped; the final values are given in Table 2.

Software Implementation

As pointed out by Mayo et al.,[48] the pairwise energy function and discrete conformational space imply that all the relevant energy data can be precomputed and stored in an energy matrix.[10] In effect, we must compute the interactions between all pairs of amino acids in the structure, allowing for all possible pairwise combinations of amino acid types and rotamer values. This calculation is done with the XPLOR program,[46] using a single command script and standard features of the program. Because of its pairwise nature and low communication requirements, this calculation can be done in parallel. We employed our Proteins @ Home distributed computing platform, which allows us to use the computers of several thousand volunteers in over 70 countries (see the list of participants at biology.polytechnique.fr/proteinsathome). Proteins @ Home is based on the Berkeley Open Infrastructure for Network Computing, BOINC.[49] The Proteins @ Home platform and project will be described in detail elsewhere.[41]

In a second stage, sequence optimization is done with the heuristic algorithm described above. A C++ program was developed, called Proteus, with core routines taken from the Optimizer program of Wernisch et al.[10] Postprocessing of the computed sequences is done with Proteus and a set of Perl scripts. The XPLOR scripts and Proteus program are available on request (thomas.simonson@polytechnique.fr). The CPU requirements of the different steps are given in Table 3.

Table 2. Empirical Corrections to Unfolded Energy (kcal/mol).

Residue type X	Original contribution	Optimized contribution	Empirical correction e_X
Ala	−2.80	−1.50	−1.30
Asp	−22.17	−15.82	−6.35
Asn	−12.85	−7.81	−5.04
Arg	−20.89	−23.35	2.46
Glu	−22.74	−17.75	−4.99
Gln	−12.02	−8.90	−3.12
His	−13.22	−11.40	−1.82
Ile	−5.63	−5.19	−0.44
Leu	−5.55	−4.63	−0.92
Lys	−16.69	−18.81	2.12
Met	−5.61	−6.99	1.38
Phe	−7.05	−7.16	0.11
Ser	−9.09	−4.14	−4.95
Tyr	−9.77	−9.94	0.17
Thr	−8.64	−4.17	−4.47
Trp	−7.60	−11.37	3.77
Val	−5.50	−2.77	−2.73

Table 3. CPU Timing on a 2.6 GHz Intel Xeon Processor.

Protein	Grb2 (1gcqB)
Length	57 amino acids
Number of interacting residue pairs	453
Total CPU time for energy calculation	203 h
CPU time per residue pair	27 min
Number of computed sequences	100,000
CPU time for sequence calculation	27 h
CPU time for 3D structure reconstruction	1 min structure

Natural Sequence Sets

For comparison to the designed sequences, we collected natural sequences from SWISSPROT. For each of the eight SH3 proteins, we retrieved homologues with at least 60% identity, giving about 12 homologous sequences per SH3 protein. A larger set including more distant homologues was then formed by the union of these smaller sets, for a total of 94 natural SH3 sequences. For the other eight test proteins, we proceeded similarly. For each protein (say, lysozyme), we retrieved homologues with at least 60% identity. We did the same for three other proteins of the same SCOP family (C-type lysozyme family, in this example), giving four small sequence sets. We then formed a large sequence set by grouping the small sets. For lysozyme, the large set has 199 homologous sequences.

Molecular Dynamics Simulations

We performed MD simulations using the AMBER force field[50] and a generalized Born (GB) solvent model. Specifically, we used the GB/HCT variant,[43,51] originally proposed by Hawkins et al.[52] The GB parameters were optimized previously for computational sidechain placement and protein mutagenesis.[39] To equilibrate the structures, 50,000 MD steps were done using a timestep of 0.1 fs and the Verlet algorithm. Next, the preequilibrated proteins were gradually heated from 50 to 300 K. After the heating, 250,000 steps of equilibration were done. We then increased the timestep to 0.5 fs and equilibrated the proteins for another 400,000 steps. During the equilibration, we rescaled the atomic velocities every 250 steps. Finally, we ran MD with no velocity scaling for 2 ns.

Results

Comparing the Designed and Native Sequences; Comparison to Earlier Work

The design approach was tested on 16 globular proteins, listed in Table 1 and shown in Figure 1. The test set includes α-helical proteins (protein A, homedomain), β-proteins (lysozyme), and mixed, α, β proteins. Four SH3 proteins (uppermost in Table 1) were used as a "learning set" to optimize the empirical energy corrections, e_X (see Methods). The optimized values (see Table 2) were then used throughout this study.

Figure 2 shows two typical ΔG_{fold} spectra, corresponding to ~450,000 sequences computed for Grb2 and c-Crk. The spectra are very roughly Gaussian, with an energy range of ~60 kcal/mol. The location of the best BLOSUM62 scores (using the native sequence as a target) are indicated. The best BLOSUM scores do not correspond to the lowest energy sequences, but are distributed over the energy spectrum, especially the region of maximal sequence occurrence (the peak of the spectrum). This is not surprising, since natural sequences are usually not optimized by evolution to maximize their stability.[54] Rather, one can often increase the stability of a protein by exchanging hydrophilic core residues for hydrophobic ones.[55,56]

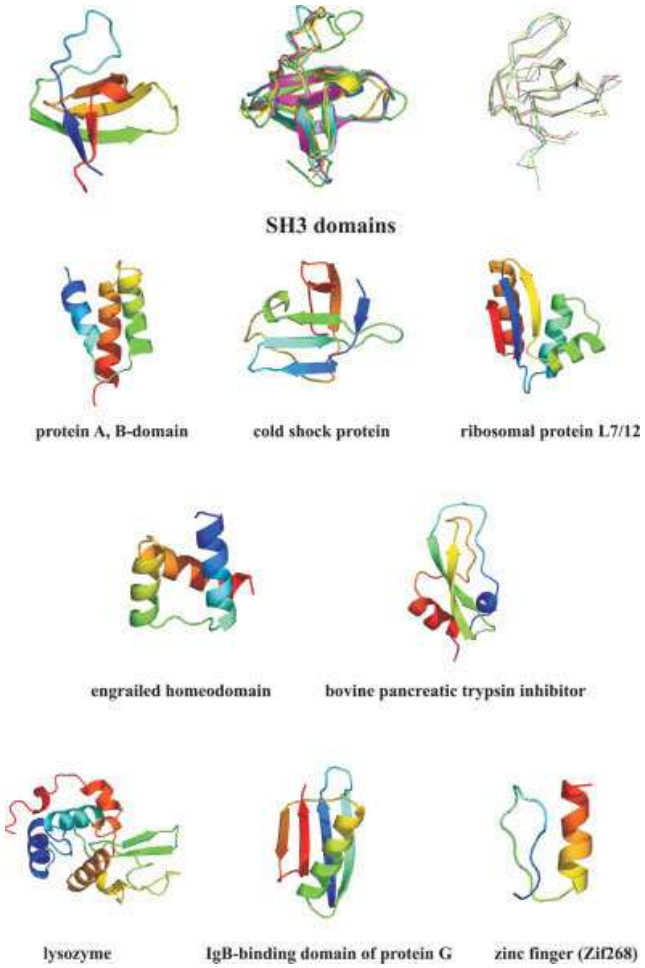


Figure 1. The test set of proteins. The eight SH3 domains (upper line) include Grb2 (left), four proteins (including Grb2) used as a learning set for the unfolded energy correction (center), and four other SH3 domains (right; shown as a structural alignment that also includes Grb2). The other eight proteins are below. Figure produced with Pymol.[53] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

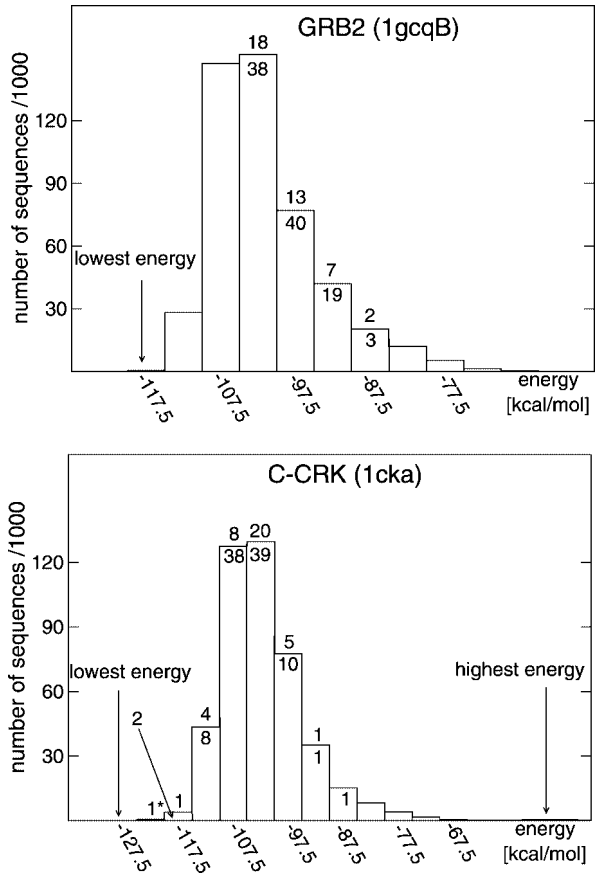


Figure 2. Histograms of Grb2 and c-Crk folding energies (ΔG_{fold}) for the computed sequences. Numbers above (within) bins indicate the number of sequences that are in the top 40 (top 100) scoring sequences (using the Blosum62 similarity matrix, with the native sequence as the target).

Table 4 gives the identity scores for the 16 proteins. "Reduced" identities are also given; they are computed with the nonmutating residues excluded (Cys, Gly, and Pro), leading to lower identities. If we consider the complete set of sequences for the eight SH3 proteins (~450,000 each), the highest average identity is for Vav (1gcqC), with 42.5% and the lowest is for the HCK kinase, with 23.5%. The four proteins of the "learning set" give an average identity of 34.5%; the remaining SH3 domains give 30.7%. The other eight proteins give an average identity of 30.4%. The averages over the low energy sequences (sequences with the highest folding free energy) are slightly higher: 35.0% for the SH3 domains and 31.8% for the other eight proteins (Table 4). If we consider only the best scoring sequences within each data set, we obviously

Table 4. Identity Scores of Computed Sequences.

PDB code	Protein name	Length (reduced)	Full identity			Reduced identity		
			Mean[a]	Low energy[b]	Best[c]	Mean	Low energy	Best
1gcqB	Grb2	57 (46)	37.3	37.2	49.5	22.4	22.1	37.5
1gcqC	Vav	69 (52)	42.5	45.6	55.0	23.7	27.8	40.2
1cka	c-Crk	56 (48)	34.5	39.8	52.1	23.6	29.8	44.2
1shg	Alpha-spectrin	57 (53)	23.6	26.9	37.6	17.8	21.4	32.9
1abo	ABL kinase	58 (49)	35.9	37.7	48.5	24.1	26.3	39.1
1ad5	HCK kinase	58 (54)	23.5	22.6	38.4	17.8	16.9	33.8
1csk	Csk	56 (46)	37.1	37.3	48.5	23.5	23.6	37.4
1fmk	c-Src	60 (54)	26.3	32.4	43.5	18.2	24.9	37.2
1lz1	Lysozyme	130 (109)	33.2	33.8	43.5	20.3	21.0	32.6
4pti	BPTI	58 (42)	38.5	36.5	50.7	15.1	12.3	31.9
1ctf	L7/12	68 (61)	32.5	31.7	44.3	24.7	23.9	41.5
1enh	Homeodomain	54 (52)	18.2	18.9	32.1	15.0	15.8	29.5
1pgb	Protein G	56 (52)	29.2	31.3	43.4	23.7	26.0	39.0
1zaa	Zif268	28 (24)	37.3	43.6	51.7	26.8	34.2	43.6
1c9o	Cold shock	66 (55)	32.4	33.0	46.1	18.8	19.6	35.4
1bdd	Protein A	53 (49)	22.1	25.8	35.0	15.7	19.7	29.6
	Average		33.2	33.4	45.0	20.7	22.8	36.6

Identity (%) with respect to the native sequence. Reduced identity takes into account only residues allowed to mutate (all except Cys, Gly, and Pro).

[a] Average over all 450,000 computed sequences.

[b] Average over the 100 lowest energy sequences.

[c] Average over the 40 highest-identity sequences.

obtain much higher identity scores: 46% for the eight SH3 proteins and 43.4% for the other eight proteins. If the native sequence is known ahead of time (the most common situation by far), these very high-scoring sequences can be identified and used. Collections of designed sequences obtained for each protein are available as Supplementary Material.

The above identity rates compare favorably to several previous studies of whole protein redesign. Larson et al. recently applied an expensive, flexible backbone approach to 253 different protein families.[24] Average identity scores between 26% and 33% were obtained, several points below our typical scores. For the Kunitz_BPTI family, an average identity of 26% was given[24]; we obtain a mean value of 38.5% for 4pti (Table 4).

In 2000, Baker and coworkers obtained an average sequence identity between computed and experimental sequences of 27% for a test set of 108 proteins.[9] To obtain this result, they optimized their energy function so that the lowest energy sequences gave the highest identity scores. Further optimization of the energy function, more elaborate rotamer libraries and the use of a flexible protein backbone enabled them to improve the designed sequences considerably, so that in 2003 they obtained an average identity of 35% for nine redesigned globular proteins.[33] In 2005, they reported an average sequence identity of 37% for the redesign of 42 small globular proteins.[15] This last result is about 4% higher than our low-energy sequences, but a bit lower than our best sequences. Their method is designed to yield a small number of highly optimized sequences. In contrast, future applications to fold recognition will require sequence sets that are both realistic and sufficiently diverse.[15,22–24]

In a more restrictive design procedure, Koehl and Levitt mutated residues by exchanging existing pairs, thus constraining the overall amino acid composition to the native one.[18] Their designed sequences of 1ctf and 1tim had average identities of 36 and 16%, respectively. Here, we report an average identity of 31.7% for our lowest-energy sequences (and 44.3% for the best-scoring sequences). In 1997, Dahiyat et al. described the first completely redesigned protein, the zinc finger Zif268.[48] Their best sequence had an identity of 36% with the native protein. Here, our highest identity score is 51.7%.

In 2005, Pokala and Handel carried out full protein design for seven proteins.[25] They reported identity values for two somewhat different approaches. The best approach differs by an optimization of the van der Waals energy term and by the incorporation of a negative design criterion, which restricts the amino acid composition at the protein surface. This approach gave identity scores between 33.5 and 46.7%. The only protein common to our study is protein G. Their best method gave an average identity of 41.1%, while their other method gave 31.1%. It is not clear how much of the difference is due to the van der Waals energy and how much is due to the surface composition restraint. Our method, which does not restrain surface composition, gave 31.3% for the low energy sequences. The good results of Pokala and Handel with either method are also due to their more detailed, all-atom force field (OPLS-AA) and their expensive, generalized Born solvation model.

The method most similar to ours is that of Wodak and coworkers, applied in 2002 to seven different protein folds.[36] Four of them are represented in our data set: SH3 domains, the homeobox, protein G, and the cold shock protein. Despite their using

a more complex, all-atom force field and a more complex rotamer library,[57] our results are significantly improved. Thus, for 11 SH3 domains, they reported an average identity score of 23.9%, compared to 34.9% obtained here for eight SH3 domains. We average over our 40 lowest-energy sequences, which is comparable to the averaging method employed in.[36] We also have notably improved scores for the cold shock protein, the homeodomain, and protein G (Table 4).

The Computed Sequences are Similar to Experimental Sequences

To compare the designed sequences to natural sequences in more detail, we created two sets of natural SH3 sequences. First, for each of our eight SH3 proteins, we retrieved from SWISSPROT sequences that share at least 60% identity with the query sequence. The second set of natural sequences includes 94 entries and is simply the union of the smaller SH3 sets collected above. Figure 3 compares the identity and BLOSUM scores obtained with the natural and computed sequences; see also Table 5 and Supplementary Material. We applied the following weighting scheme to the sequences. The variability of each amino acid within the larger set of natural sequences was computed. If the frequency of the prevailing residue exceeded 80%, a weighting factor of 1 was assigned. A frequency between 50 and 80% led to a weight of 0.75, and a frequency of less than 50% led to a weight of 0.5. In most cases, the computed scores (Fig. 3) overlap with the range of the large set of natural sequences. For four of the SH3 domains (1gcqC,1cka, 1abo, and 1csk), the weighted average BLOSUM62 score of the computed sequences actually exceeds or is very close to the corresponding value of the natural set. Thus, in most cases, the computed sequences behave like moderately-distant homologues of the native

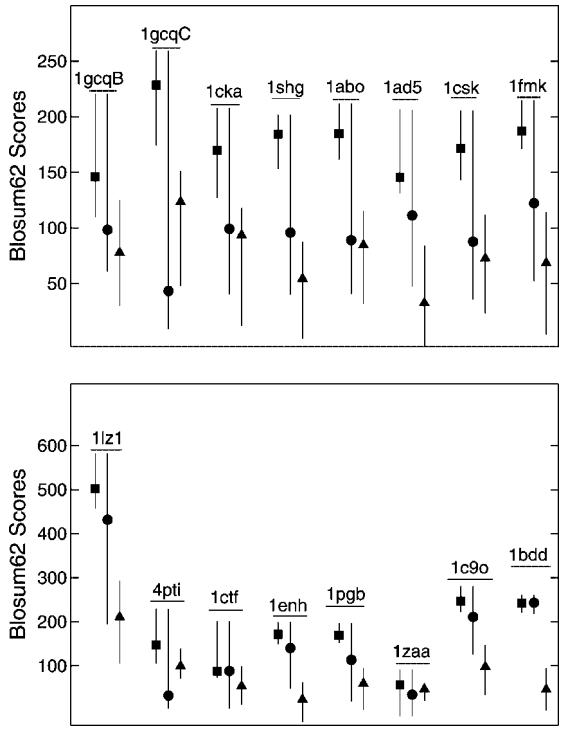


Figure 3. Comparison of native and computed BLOSUM62 scores. Proteins are identified by their PDB code. For each protein, three vertical lines represent the range of scores within the small natural set (left), the large natural set (middle), and the computed set (right). The average score in each set is also shown (as a square, a dot, a triangle).

Table 5. Comparison of BLOSUM62 Scores Between Natural and Computed Sequences.

PDB code	Unweighted scores				Weighted scores			
	Natural	Total[a]	Low energy[b]	Best[c]	Natural	Total	Low energy	Best
1gcqB	114.1	112.8	102.1	161.1	98.2	84.8	77.9	119.4
1gcqC	29.7	149.8	163.2	204.5	43.2	111.8	123.5	145.5
1cka	113.0	105.1	129.5	159.4	99.1	77.6	93.6	111.8
1shg	111.6	51.1	64.4	101.2	95.9	41.3	54.3	79.3
1abo	104.0	103.5	109.8	145.4	89.0	80.4	84.7	106.2
1ad5	134.0	45.0	33.7	89.5	111.3	39.8	32.5	69.3
1csk	98.0	99.8	99.5	147.5	87.7	70.6	72.7	105.2
1fmk	151.3	61.9	82.5	127.2	122.2	52.6	68.6	101.2
1lz1	487.9	236.6	244.8	324.3	431.1	205.2	211.0	278.2
4pti	37.0	138.8	131.1	181.2	31.6	104.1	99.4	131.1
1ctf	71.5	91.5	86.0	142.3	65.1	58.4	53.9	92.35
1enh	156.6	25.3	28.7	58.6	139.6	19.0	23.7	44.62
1pgb	154.1	61.6	73.9	115.8	113.5	50.4	59.8	86.24
1zaa	24.8	61.14	73.0	82.0	22.7	39.9	47.1	52.1
1c9o	232.1	97.1	98.6	152.3	160.2	93.7	96.7	139.7
1bdd	246.2	46.5	45.5	81.9	238.5	47.3	46.4	81.2

BLOSUM62 scores with respect to the native sequence. Weighting scheme is described in the main text.
[a] Average over all 450,000 computed sequences.
[b] Average over the 100 lowest energy sequences.
[c] Average over the 100 highest-identity sequences.

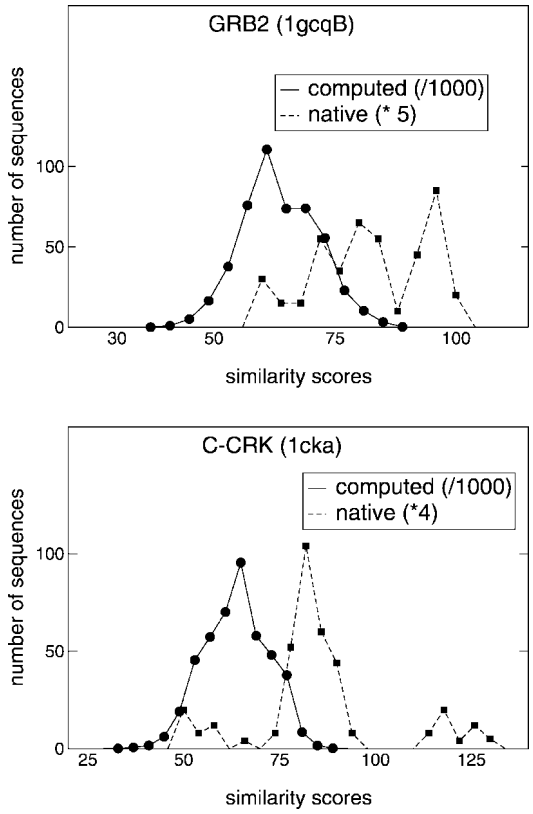


Figure 4. Experimental and computed similarity scores for two SH3 proteins: Grb2 and c-Crk. Dashed: Blosum62 scores for 94 natural SH3 domains, compared to the native sequence. Black: scores for the computed sequences.

sequence. The same trend is mostly seen for the other eight proteins. Only for 1bdd are the computed sequences distinctly below the natural set.

Figure 4 further illustrates the designed sequences for the four SH3 proteins of our learning set. The computed sequences are compared to a sequence profile obtained from the large set of 94 natural SH3 sequences. Similarity scores were computed using the formula:[36]

$$s = \sum_i \sum_a f_{ia} S(x_i, a). \tag{3}$$

Here, x_i is the amino acid type at position i in the native sequence; a is one of the 20 amino acid types, f_{ia} is the frequency (between 0 and 1) of amino acid type a at position i in the set of natural sequences; $S(x_i, a)$ is the BLOSUM62 similarity score; the first sum is over the native sequence; the second sum is over the amino acid types. The values of s are given on the horizontal axis of Figure 4 and the number of scores are given on the vertical axis. The similarity scores of the designed sequences fall within the range of the native scores in all four cases. The similarity scores of the designed sequences span the range 30–90, whereas the native scores span the range 50–125. In three of the four cases (Grb2, c-Crk, and alpha-spectrin), the designed sequences overlap with the lower part of the spectrum of natural sequences. For Vav, the designed sequences overlap with the upper part of the spectrum. Overall, the similarity scores are consistent with results presented in Figure 3. As with the BLOSUM scores, we find that the similarity scores of the designed sequences overlap with the scores of natural sequences of other SH3 proteins.

Finally, BLAST searches of the SWISSPROT databank were done using the best-scoring BLOSUM62 sequences of the eight SH3 proteins. In all eight cases, a series of native SH3 proteins were retrieved, which are shown in Figures 5 and 6 for Grb2 and c-crk, respectively. Encouragingly, the native sequence was retrieved in all eight cases. No high-scoring false-positives were obtained; e.g., all the sequences with an E-value below 0.001 were indeed SH3 domains. Conversely, when an unrelated protein of a comparable size, BPTI, is used as a query, no SH3 sequences are retrieved among the top 1000 sequences returned by BLAST; only sequences from the BPTI_Kunitz SCOP family are among the top 1000 sequences. The same is true if a high-scoring designed BPTI sequence is used as a query.

Analysis of Grb2 and c-Crk Sequences

Figure 5 illustrates the designed sequences for Grb2. Three blocks of sequences are shown: the 25 best-scoring designed sequences (top); 25 homologues retrieved from SWISSPROT with the native sequence as a query (middle); 25 homologues retrieved from SWISSPROT with the best-scoring designed sequence as a query (bottom). The 3D structures shown above the sequences depict the original protein and the best BLOSUM62-scoring sequence. Sequences are color-coded as follows: blue, DENQ; red, HKR; yellow, FWY; pink, ST; orange, ILMV; green, ACGP.

A long loop, typical of SH3 domains, extends from Ala5 to Gly22. In all the natural sequences, the L_6F_7 motif at the beginning of the loop is either maintained or mutated to one of the following, homologous motifs: I_6Y_7, I_6Y_7, or V_6Y_7. L_6 and F_7 are surface residues, whose sidechains point towards solvent. In the designed sequences more hydrophilic residues are found: K_6A_7 or K_6N_7. The next stretch of residues, predominantly DF(D/E), is closely mimicked by the designed motif, EFN, which occurs in nine out of 25 computed sequences. For the other 16 sequences, we find EFA or KFE. The remaining part of the loop shows a similar level of conservation. At positions 16 and 17, the conserved motif $(E/D)_{16}\,L_{17}$ is usually changed to $Y_{16}L_{17}$ and rarely to $A_{16}L_{17}$ in the designed sequences. Inspection of the native sequences of Vav (1gcqC) shows an aromatic residue in the same position as in the designed sequences. This aromatic residue interacts with other hydrophobic residues at the end of β strands 3 and 4. Therefore, one can assume that this mutation will not affect the overall stability of the protein.

The core of the Grb2 structure is a five-stranded beta sheet. In this region, designed and natural sequences agree well. The first strand includes either $Y_2V_3Q_4$, $K_2V_3Q_4$, or $K_2V_3R_4$. The designed sequences start with an ionic amino acid followed by either V_2 or an aromatic residue $(Y/F)_3$, then either an ionized residue (E/K) or a hydrophobic one (W/V). In one case, the natural $K_2V_3Q_4$ motif is reproduced. The second strand goes from G_{22} to D_{29}. The natural $(D/E)_{23}$ is usually maintained. The weakly conserved native $F_{24}I_{25}$ is mostly changed to $H_{24}I_{25}$. The solvent-exposed H_{26} is replaced in the designed sequences by a charged E_{26} or K_{26}. Finally, the

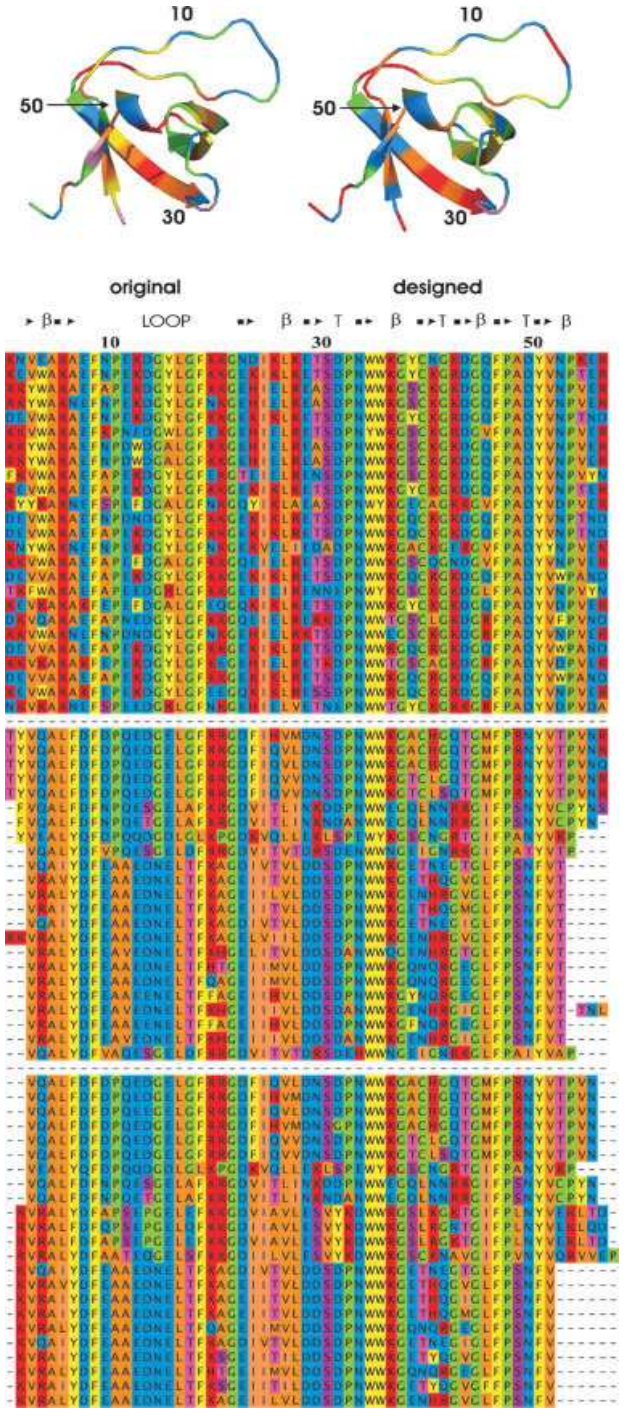


Figure 5. Grb2 sequences. Upper group: 25 high-scoring computed sequences. Middle group: 23 natural sequences obtained from SWISSPROT, with the native sequence as query. Bottom group: 23 natural sequences from SWISSPROT using a computed sequence as query. The colors distinguish six amino acid groups: {DENQ}, {HKR}, {FYW}, {ACGP}, {ST}, and {ILMV}. Secondary structure and residue numbers are shown, along with the 3D structure (colored according to the native and a designed sequence).

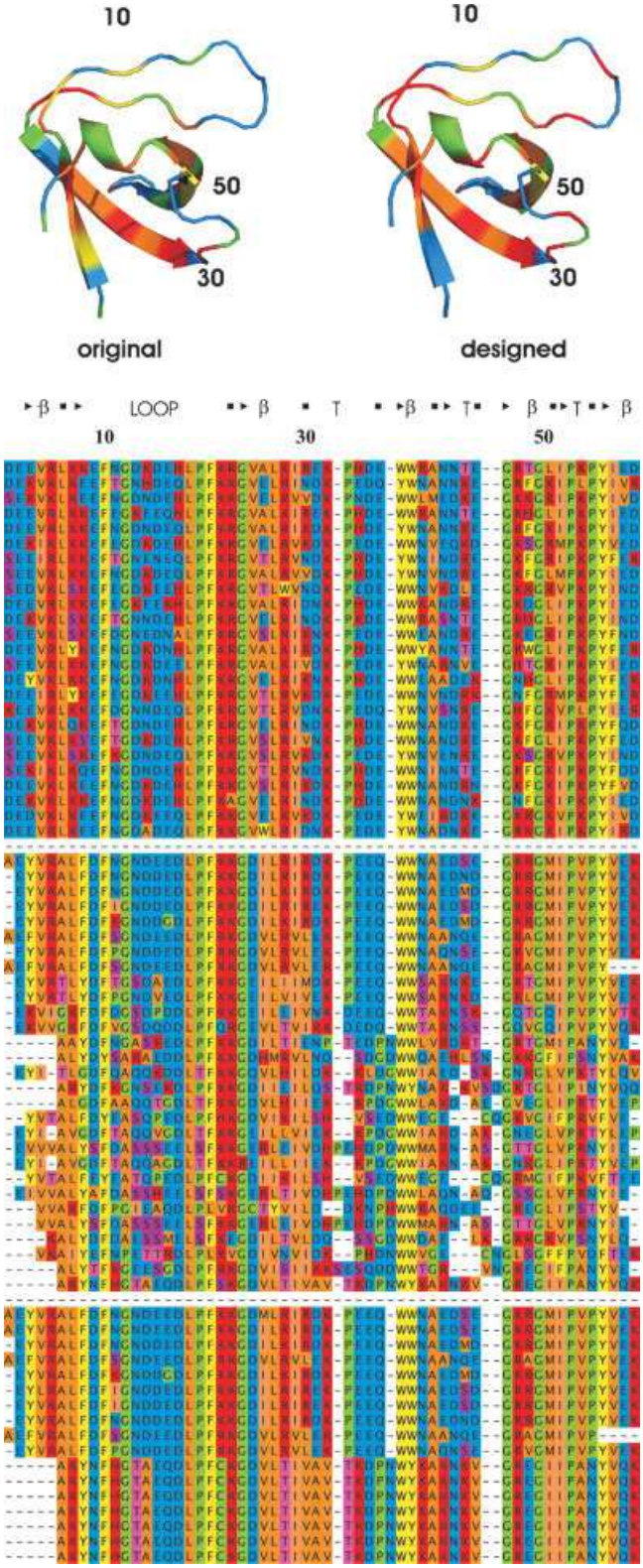


Figure 6. 1CKA sequences. Same representation as in Figure 5.

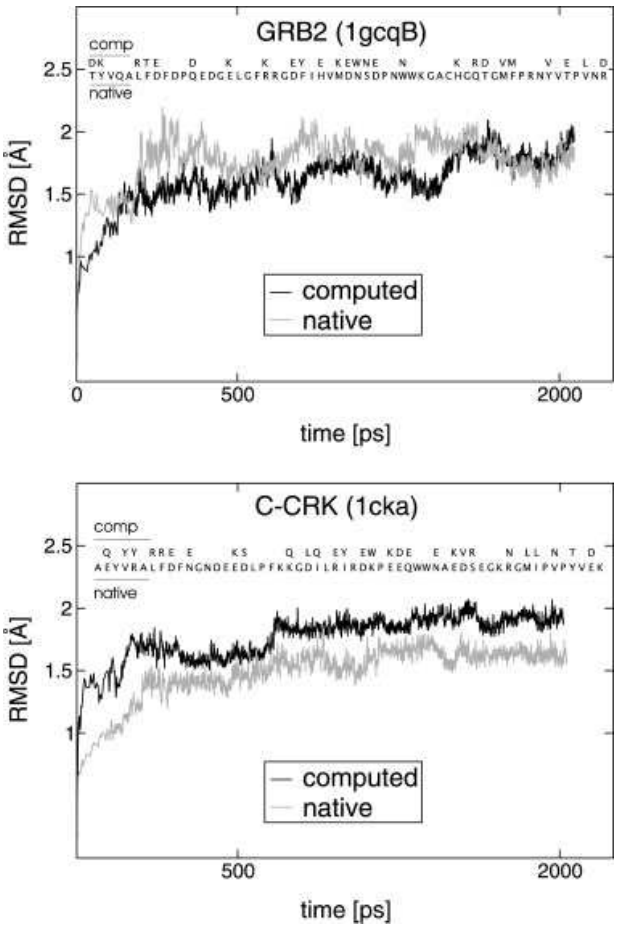


Figure 7. Molecular dynamics simulations of two native and designed proteins. RMS deviation (Å) from the initial, experimental, native structure. For each protein, the native protein and one designed variant were simulated. The native and computed sequences are shown as insets (only mutated positions in the computed sequence are shown). The AMBER force field and a generalized Born solvent were used.

native, hydrophobic-hydrophobic-ionic motif is usually changed to a hydrophobic-ionic-ionic pattern. The loop that connects strands two and three is accurately reproduced. At the beginning of the third strand, the native $W_{35}W_{36}K_{37}$ pattern is maintained. After the third strand, the prevalence of $K_{43}D_{44}$ in the designed sequences is due to a new salt bridge. Finally, the rest of the Grb2 sequence is fairly well-reproduced by the design.

C-Crk (Fig. 6) has a longer first beta strand than Grb2. The predominant native motif, $E_2Y_3V_4R_5$, is reproduced in three out of four residues ($E_2V_4R(/K)_5$), whereas in postion three we find mostly E or D. In one case, the complete native sequence is recapitulated. The following, long loop (A_2 to G_{23}) shows a strong overlap between designed and native sequences, but the L_7F_8 pattern at the beginning is not reproduced. The subsequent beta strand, from G_{23} to D_{30}, is a mixture of hydrophilic and hydrophobic amino acids in both sets of sequences. The native $W_{38}W_{39}$ motif is conserved in the designed sequences.

The Stability of the Designed Proteins is Supported by Molecular Dynamics

To further test the stability of the designed sequences we performed molecular dynamics simulations (MD) for the four proteins of the "learning set:" Grb2, Vav, c-Crk and spectrin. In each case, we considered both the native protein and a high-scoring sequence. For reasons of efficiency, we used an implicit solvent model, as for the design calculations. However, both the force field and the solvent model used here were of a higher quality than the ones used for the design calculations. Indeed, we used the all-atom, AMBER force field,[50] instead of the "polar hydrogen," Charmm19 force field[38] used above. Instead of the CASA solvent model, we used a recent, high-quality, generalized Born model (GB), which is known to yield good quality protein structures in MD simulations.[39,43,58] MD simulations were run for 2 ns in each case.

The deviations of the Grb2 and c-Crk structures from the experimental, crystal structures are shown in Figure 7 as a function of time. Data for Vav and spectrin are similar. For the four native sequences, the MD structures agree very well with the crystal structures, with rms deviations of about 1.5 or 2 Å after 2 ns. This is comparable to other proteins studied with GB,[58,59] and with explicit solvent treatments. For the designed sequences, the quality of the MD structures is comparable (Grb2, Vav) or very slightly worse (c-Crk, spectrin), with rms deviations of about 1.7–2.2 Å after 2 ns of MD. These low values suggest that the designed sequences are stable, but prefer a slightly different backbone conformation.

Conclusions

The overall design procedure implemented here combines several well-established ingredients: a molecular mechanics energy function, implicit solvent, a fixed backbone, and sidechain rotamers.[9,31,36,48] Sequences are selected based on their folding free energy, using a tripeptide model of the unfolded state. At a more detailed level, compared to previous studies of whole protein redesign, there are significant differences in one or more of the ingredients in each case. In general, our procedure uses the simplest ingredients: a "polar hydrogen" force field, a surface area solvent model, a fixed protein backbone, a heuristic method[10] for exploring sequence and rotamer space. Nevertheless, our results are comparable to other recent studies. The good results obtained here can be attributed to our earlier, complete reparameterization of the CASA solvent model[39] and a careful optimization of other model parameters. In particular, an empirical correction to the unfolded state energy was parameterized.

Thus, the present study extends our knowledge of computational protein design and its robustness or sensitivity to model details. We also tested the use of large-scale volunteer computing for this application, with our BOINC-based, Proteins @ Home platform. For eight SH3 domains and a heterogeneous set of eight other proteins, we obtained designed sequences with a native-like character. In all cases tested, the best designed sequences allowed us to retrieve the native sequence from the SWISSPROT database (with no false positives), suggesting that the method has the potential to become a fold recognition method.

The method has several obvious limitations, some of which are shared by competing implementations. By selecting for low folding energies, we could over-optimize stability, compared to natural evolution. By focussing on sequences with a high similarity to native, we could miss sequences that have a low sequence homology but a high structural homology. Nevertheless, the overall performance of the method appears encouraging. Many additional questions remain open. The amount of diversity within the computed sequences is promising but needs more investigation, for example, as well as their sensitivity to the fixed backbone assumption. In the future, we plan to implement a new, residue-pairwise, generalized Born solvent,[60] and to apply the method to problems of fold recognition.[22]

Acknowledgments

The authors thank the many volunteers who have participated in the Proteins at Home project and contributed computer cycles used in this work. We thank the BOINC development community for testing the alpha version of the Proteins at Home platform. They thank Christine Bathelt, Alexey Aleksandrov, and Najette Amara for discussions.

References

1. Drexler, K. Proc Natl Acad Sci USA 1981, 78, 5275.
2. Eisenberg, D. Nature 1982, 295, 99.
3. Pabo, C. Nature 1983, 301, 200.
4. Ponder, J.; Richards, F. M. J Mol Biol 1988, 193, 775.
5. Hellinga, H.; Richards, F. Proc Natl Acad Sci USA 1994, 91, 5803.
6. Dahiyat, B.; Mayo, S. Prot Sci 1996, 5, 895.
7. Harbury, P. B.; Plecs, J. J.; Tidor, B.; Alber, T.; Kim, P. S. Science 1998, 1998, 1462.
8. Desjarlais, J.; Handel, T. J Mol Biol 1999, 289, 305.
9. Kuhlman, B.; Baker, D. Proc Natl Acad Sci USA 2000, 97, 10383.
10. Wernisch, L.; Héry, S.; Wodak, S. J Mol Biol 2000, 301, 713.
11. Kuhlman, B.; Dantas, G.; Ireton, G.; Varani, G.; Stoddard, B.; Baker, D. Science 2003, 302, 1364.
12. Dwyer, M.; Looger, L.; Hellinga, H. Science 2004, 304, 1967.
13. Havranek, J.; Harbury, P. Nat Struct Biol 2003, 10, 45.
14. Ventura, S.; Serrano, L. Proteins 2004, 56, 1.
15. Saunders, C.; Baker, D. J Mol Biol 2005, 346, 631.
16. Wollacott, A. M.; Zanghellini, A.; Murphy, P.; Baker, D. Prot Sci 2007, 16, 165.
17. Koehl, P.; Levitt, M. J Mol Biol 1999, 293, 1161.
18. Koehl, P.; Levitt, M. J Mol Biol 1999, 293, 1183.
19. Koehl, P.; Levitt, M. Proc Natl Acad Sci USA 1999, 96, 12524.
20. Pokala, N.; Handel, T. Prot Sci 2004, 13, 925.
21. Chowdry, A. B.; Reynolds, K. A.; Hanes, M. S.; Voorhies, M.; Pokala, N.; Handel, T. J Comput Chem 2007, 28, 2378.
22. Larson, S.; Garg, A.; Desjarlais, J.; Pande, V. Proteins 2003, 51, 390.
23. Larson, S.; Pande, V. J Mol Biol 2003, 332, 275.
24. Larson, S.; England, J. E.; Desjarlais, J.; Pande, V. Prot Sci 2002, 11, 2804.
25. Pokala, N.; Handel, T. M. J Mol Biol 2005, 347, 203.
26. Raha, K.; Wollacott, A. M.; Italia, M. J.; Desjarlais, J. R. Prot Sci 2000, 9, 1106.
27. Cochran, F. V.; Wu, S. P.; Wang, W.; Nanda, V.; Saven, J. G.; Therien, M. J.; DeGrado, W. F. J Am Chem Soc 2005, 127, 1346.
28. Swift, J.; Wehbi, W. A.; Kelly, B. D.; Stowell, X. F.; Saven, J. G.; Dmochowski, I. J. J Am Chem Soc 2006, 128, 6611.
29. Kang, S. G.; Saven, J. G. Curr Opin Chem Biol 2007, 11, 329.
30. Baker, D. Phil Trans R Soc Lond 2006, 361, 459.
31. Butterfoss, G.; Kuhlman, B. Ann Rev Biophys Biomolec Struct 2006, 35, 49.
32. Guérois, R.; Lopez de la Paz, M. Eds. Protein Design: Methods And Applications; Humana Press: New Jersey 2007.
33. Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. J Mol Biol 2003, 332, 449.
34. Koehl, P.; Levitt, M. Proc Natl Acad Sci USA 2002, 99, 1280.
35. Koehl, P.; Levitt, M. Proc Natl Acad Sci USA 2002, 99, 691.
36. Jaramillo, A.; Wernisch, L.; Héry, S.; Wodak, S. Proc Natl Acad Sci USA 2002, 99, 13554.
37. Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. J Biomol Struct Dyn 1991, 8, 1267.
38. Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. J Comp Chem 1983, 4, 187.
39. Lopes, A.; Aleksandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T. Proteins 2007, 67, 853.
40. Liang, S.; Grishin, N. Proteins 2004, 54, 271.
41. Simonson, T.; Mignon, D.; Schmidt am Busch, M.; Lopes, A.; Bathelt, C. In Distributed and Grid Computing—Science Made Transparent for Everyone. Principles, Applications and Supporting Communities; Tektum Publishers: Berlin, 2007.
42. Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, J. J.; Chothia, C.; Murzin, A. G. Nucl Acids Res 2004, 32, D226.
43. Onufriev, A.; Bashford, D.; Case, D. J Phys Chem B 2000, 104, 3712.
44. Fraternali, F.; van Gunsteren, W. J Mol Biol 1996, 256, 939.
45. Lee, B.; Richards, F. J Mol Biol 1971, 55, 379.
46. Brünger, A. T. X-plor version 3.1, A System for X-Ray Crystallography and NMR; Yale University Press: New Haven, 1992.
47. Street, A.; Mayo, S. Fold Desig 1998, 3, 253.
48. Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Prot Sci 1997, 6, 1333.
49. Anderson, D. P. Boinc: A System for Public-Resource Computing and Storage. In 5th IEEE/ACM International Workshop on Grid Computing; IEEE Computer Society Press; USA; 2004.
50. Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. J Am Chem Soc 1995, 117, 5179.
51. Moulinier, L.; Case, D.; Simonson, T. Acta Cryst D 2003, 59, 2094.
52. Hawkins, G.; Cramer, C.; Truhlar, D. Chem Phys Lett 1995, 246, 122.
53. DeLano, W. L. The PyMOL Molecular Graphics System; DeLano Scientific: San Carlos, CA, USA, 2002.
54. DePristo, M. A.; Weinreich, D. M.; Hartl, D. L. Nature Rev Genet 2005, 6, 678.
55. Stites, W.; Gittis, A.; Lattman, E.; Shortle, D. J Mol Biol 1991, 221, 7.
56. Shortle, D. Curr Opin Struct Biol 1993, 3, 66.
57. Dunbrack, R.; Karplus, M. J Mol Biol 1993, 230, 543.
58. Simonson, T.; Carlsson, J.; Case, D. A. J Am Chem Soc 2004, 126, 4167.
59. Feig, M.; Brooks, C. L., III. Curr Opin Struct Biol 2004, 14, 217.
60. Archontis, G.; Simonson, T. J Phys Chem B 2005, 109, 22667.

Bibliographie

[1] L. Ador, A. Camasses, P. Erbs, J. Cavarelli, D. Moras, J. Gangloff, and G. Eriani. Active site mapping of yeast aspartyl-tRNA synthetase by *in vivo* selection of enzyme mutations lethal for cell growth. *J. Mol. Biol.*, 288 :231–242, 1999.

[2] F. Agou, S. Quevillon, P. Kerjan, and M. Mirande. Switching the amino acid specificity of an aminoacyl-trna synthetase. *Biochemisry*, 37 :11309–11314, 1998.

[3] E. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native strutures. *Proc. Natl. Acad. Sci. USA*, 96 :11305–11310, 1999.

[4] G. Archontis and T. Simonson. A residue-pairwise Generalized Born scheme suitable for protein design calculations. *J. Phys. Chem. B*, 109 :22667–22673, 2005.

[5] G. Archontis, T. Simonson, and M. Karplus. Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase. *J. Mol. Biol.*, 306 :307–327, 2001.

[6] G. Archontis, T. Simonson, D. Moras, and M. Karplus. Specific amino acid recognition by aspartyl-trna synthetase studied by free energy simulations. *J. Mol. Biol.*, 275 :823–846, 1998.

[7] J. Arnez and D. Moras. *Aminoacyl-tRNA synthetases-tRNA recognition.* edition K. nagai I. mattaj, 1994.

[8] D. Bashford and M. Karplus. The pK_a's of ionizable groups in proteins : atomic detail from a continuum electrostatic model. *Biochemistry*, 29 :10219–10225, 1990.

[9] A.A. Beaudry and G.F. Joyce. Directed evolution of an rna enzyme. *Science*, 257 :635–641, 1992.

[10] H. Belrhali, Yaremchuk A, Tukalo M, Berthet-Colominas C, Rasmussen B, Bosecke P, Diat O, and Cusack S. The structural basis for seryl-adenylate and ap4a synthesis by seryl-trna synthetase. *Structure*, 3 :341–352, 1995.

[11] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28 :235–242, 2000.

[12] C. Berthet-Colominas, L. Seignovert, M. Hartlein, M. Grotli, S. Cusack, and R. Leberman. The crystal structure of asparaginyl-trna synthetase from thermus thermophilus and its complexes with atp and asparaginyl-adenylate : the mechanism of discrimination between asparagine and aspartic acid. *EMBO J.*, 17 :2947–2960, 1998.

[13] F.E. Boas and P.B. Harbury. Potential energy functions for protein design. *Curr. Opin. Struct. Biol.*, 17 :199–204, 2007.

[14] D.N. Bolon, J.S. Marcus, S.A. Ross, and S.L. Mayo. Prudent modeling of core polar residues in computational protein design. *J. Mol. Biol.*, 329 :611–622, 2003.

[15] S. Boresch and M. Karplus. The meaning of component analysis : decomposition of the free energy in terms of specific interactions. *J. Mol. Biol.*, 254 :801–807, 1995.

[16] P. Brick, T.N. Bhat, and D.M. Blow. Structure of tyrosyl-tRNA synthetase refined at 2.3 å resolution. Interaction of the enzyme with the tyrosyl adenylate intermediate. *J. Mol. Biol.*, 208 :83–98, 1989.

[17] B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus. Charmm : a program for macromolecular energy, minimization, and molecular dynamics calculations. *J. Comp. Chem.*, 4 :187–217, 1983.

[18] C.L. Brooks, M. Karplus, and M. Pettitt. Proteins : a theoretical perspective of dynamics, structure and thermodynamics. *Adv. Chem. Phys.*, 71 :1–259, 1987.

[19] J.R. Brown and W.F. Doolittle. Root of the universal tree of life based on ancient aminoacyl-trna synthetase gene duplications. *Proc. Natl. Acad. Sci. USA*, 92 :2441–2445, 1995.

[20] A. T. Brünger. X-PLOR *version 3.1, A System for X-ray crystallography and NMR.* Yale University Press, New Haven, 1992.

[21] S. Brunie, C. Zelwer, and J.L. Risler. Crystallographic study at 2.5 å resolution of methionyl-tRNA synthetase from *Escherichia coli* with ATP. *J. Mol. Biol.*, 216 :411–424, 1990.

[22] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes. Cfunnels, pathways, and the energy landscape of protein folding : a synthesis. *Proteins*, 21 :167–195, 1995.

[23] J.J. Burbaum, R.M. Starzyk, and P. Schimmel. Understanding structural relationships in proteins of unsolved three-dimensional structure. *Proteins*, 7 :99–111, 1990.

[24] N. Calimet, Michael Schaefer, and Thomas Simonson. Protein molecular dynamics with the Generalized Born/ACE solvent model. *Proteins*, 45 :144–158, 2001.

[25] D.A. Case, D.A. Pearlman, J.C. Caldwell, T.E. Cheatham III, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, V. Tsui, R.J. Radmer, Y. Duan, J. Pitera, I. Massova, G.L. Seibel, U.C. Singh, P.K. Weiner, and P.A. Kollman. *AMBER 6.* (University of California, San Francisco), 1999.

[26] J. Cavarelli, G. Eriani, B. Rees, M. Ruff, M. Boeglin, A. Mitschler, F. Martin, J. Gangloff, J.C. Thierry, and D. Moras. The active site of yeast aspartyl-tRNA synthetase : structural and functional aspects of the aminoacylation reaction. *EMBO J.*, 13 :327–337, 1994.

[27] J. Cavarelli and D. Moras. Recognition of trnas by aminoacyl-trna synthetases. *FASEB J.*, 7 :79–86, 1993.

[28] J. Cavarelli, B. Rees, M. Ruff, J.C. Thierry, and D. Moras. Yeast tRNA(Asp) recognition by its cognate class II aminoacyl-tRNA synthetase. *Nature*, 362 :181–184, 1993.

[29] J. Cavarelli, B. Rees, J.C. Thierry, and D. Moras. Yeast aspartyl-trna synthetase : a structural view of the aminoacylation reaction. *Biochimie*, 75 :1117–1123, 1993.

[30] F. Chapeville, F. Lipmann, G. Von Ehrenstein, B. Weisblum, and R.W.J. Benzers. On the role of soluble ribonucleic acid in coding for amino acids. *Proc. Natl. Acad. Sci. USA*, 48 :1086–1092, 1962.

[31] B.S. Chevalier, T. Kortemme, M.S. Chadsey, D. Baker, R.J. Monnat, and B.L. Stoddard. Design, activity, and structure of a highly specific artificial endonuclease. *Moll. Cell*, 10 :895–905, 2002.

[32] J.W. Chin, T.A. Cropp, J.C. Anderson, M. Mukherji, Z. Zhang, and P.G. Schultz. An expanded eukaryotic genetic code. *Science*, 301 :964–967, 2003.

[33] F. Chiti, N. Taddei, P. Webster, D. Hamada, T. Fiaschi, G. Ramponi, and C.M. Dobson. Acceleration of the folding of acylphophatase by stabilization of local secondary structure. *Nat. Struct. Biol.*, 6 :380–387, 1999.

[34] T.L. Chiu and R.A. Goldstein. How to generate improved potentials for proteins tertiary structure prediction : a lattice model study. *Proteins*, 41 :157–163, 2000.

[35] M . Clamp, J. Cuff, S.M. Searle, and G.J. Barton. The jalview java alignment editor. *Bioinformatics*, 20 :426–427, 2004.

[36] F. Crick. The packing of alpha-helicies : simple coiled-coils. *Acta. Crystallogr.*, 6 :689–697, 1953.

[37] F.H. Crick. The origin of the genetic code. *J. Mol. Biol.*, 38 :367–379, 1968.

[38] T.A. Cropp and P.G. Schultz. An expanding genetic code. *Trends. Genet.*, 20 :625–630, 2004.

[39] S. Cusack, C. Berthet-Colominas, M. Hartlein, N. Nassar, and N. Leberman. Sequence, structure and evolutionary relationships between class 2 aminoacyl-trna synthetases. *Nucleic Acids Research*, 19 :3489–3498, 1991.

[40] S. Cusack, M. Hartlein, and N. Leberman. A second class of synthetase structure revealed by Xray analysis of *Escherichia coli* seryl-tRNA synthetase at 2.5 å˙ *Nature*, 347 :249–255, 1990.

[41] B.I. Dahiyat. *In silico* design for protein stabilization. *Curr. Opin. Biotechnol.*, 10 :387–390, 1999.

[42] B.I. Dahiyat, D.B. Gordon, and S.L. Mayo. Automated design of the surface positions of protein helices. *Protein Science*, 6 :1333–1337, 1997.

[43] B.I. Dahiyat and S.L. Mayo. Protein design automation. *Protein Science*, 5 :895–903, 1996.

[44] B.I. Dahiyat and S.L. Mayo. De novo protein design : fully automated sequence selection. *Science*, 278 :82–87, 1997.

[45] B.I. Dahiyat and S.L. Mayo. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA*, 94 :10172–10177, 1997.

[46] B.I. Dahiyat, C.A. Sarisky, and S.L. Mayo. De novo protein design : towards fully automated sequence selection. *J. Mol. Biol.*, 273 :789–796, 1997.

[47] G. Dantas, B. Kuhlman, D. Callender, M. Wong, and D. Baker. A large test of computational protein design : Folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.*, 332 :449–460, 2003.

[48] L. David, R. Luo, and M. Gilson. Comparison of Generalized Born and Poisson models : energetics and dynamics of HIV protease. *J. Comput. Chem.*, 21 :295–309, 2000.

[49] M. Delarue. Partition of aminoacyl-trna synthetases in two different structural classes dating back to early metabolism : implications for the origin of the genetic code and the nature of protein sequences. *J. Mol. Evol.*, 41 :703–711, 1995.

[50] M. Delarue. An asymmetric underlying rule in the assignement of codons : Possible clue to a quick early evolution of the genetic code via successive binary choices. *RNA*, 13 :161–169, 2007.

[51] M. Delarue and D. Moras. The aminoacyl-trna synthetase family : modules at work. *Bioessays*, 15 :675–687, 1995.

[52] M. Delarue, A. Poterszman, S. Nikonov, M. Garber, D. Moras, and J.C. Thierry. Crystal structure of a procaryotic aspartyl-trna-synthetase. *EMBO J.*, 13 :3219–3229, 1995.

[53] J.R. Desjarlais and N.D. Clarke. Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.*, 8 :471–475, 1998.

[54] J.R. Desjarlais and T.M. Handel. De novo design of the hydrophobic cores of proteins. *Protein Science*, 4 :2006–2018, 1995.

[55] J.R. Desjarlais and T.M. Handel. Sidechain and backbone flexibility in protein core design. *J. Mol. Biol.*, 289 :305–318, 1999.

[56] J. Desmet, M. De Mayaer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein sidechain positioning. *Nature*, 356 :539–542, 1992.

[57] B. Dominy and C.L. Brooks III. Development of a Generalized Born model parameterization for proteins and nucleic acids. *J. Phys. Chem. B*, 103 :3765–3773, 1999.

[58] B.N. Dominy and C.L. Brooks. Identifying native-like protein structures using physics-based potentials. *J. Comput. Chem.*, 23 :147–160, 2002.

[59] V. Doring, HD. Mootz, L.A. Nangle, T.L. Hendrickson, V. de Crecy-Lagard, P. Schimmel, and P. Marliere. Enlarging the amino acid set of escherichia coli by infiltration of the valine coding pathway. *Science*, 292 :501–504, 2001.

[60] M.A. Dwyer, L. Looger, and H. Hellinga. Computational design of a biologically active enzyme. *Science*, 304 :1967–1971, 2004.

[61] D. Eisenberg and A.D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319 :199–203, 1986.

[62] A.D. Ellington and J.M. Szostak. In vitro selection of rna molecules that bind specific ligands. *Nature*, 346 :818–822, 1990.

[63] G. Eriani, J. Cavarelli, F. Martin, L. Ador, B. Rees, J.C. Thierry, J. Gangloff, and D. Moras. The class ii aminoacyl-trna synthetases and their active site : evolutionary conservation of an atp binding site. *J. Mol. Evol.*, 40 :499–508, 1995.

[64] G. Eriani, J. Cavarelli, F. Martin, G. Dirheimer, D. Moras, and J. Gangloff. Role of dimerization in yeast aspartyl-tRNA synthetase and importance of the class II invariant proline. *Proc. Natl. Acad. Sci. USA*, 90 :10816–10820, 1993.

[65] G. Eriani, M. Delarue, O. Poch, J. Gangloff, and D. Moras. Partition of aminoacyl-tRNA synthetases into two classes on the basis of two mutually exclusive sets of sequence motifs. *Nature*, 347 :203–206, 1990.

[66] G. Eriani, G. Prevost, D. Kern, P. Vincendon, G. Dirheimer, and J. Gangloff. Cytoplasmic aspartyl-trna synthetase from saccharomyces cerevisiae. study of its functional organisation by deletion analysis. *Eur. J. Biochem.*, 200 :337–343, 1991.

[67] A.R. Fersht. Editing mechanisms in protein synthesis. rejection of valine by the isoleucyl-trna synthetase. *Biochemistry*, 16 :1025–1030, 1977.

[68] A. V. Filikov, R. J. Hayes, P. Z. Luo, D. M. Stark, C. Chan, A. Kundu, and B. I. Dahiyat. Computational stabilization of human growth hormone. *Protein Science*, 278 :1452–1461, 2002.

[69] F. Fogolari, A. Brigo, and H. Molinari. Protocol for mm/pbsa molecular dynamics simulations of proteins. *Biophys. J.*, 85 :159–166, 2003.

[70] F. Fraternali and W.F. van Gunsteren. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *J. Mol. Biol.*, 256 :939–948, 1996.

[71] X. Fu, J.R. Apgar, and A.E. Keating. Modeling backbone flexibility to achieve sequence diversity : the design of novel α-helical ligands for Bcl-x$_L$. *J. Mol. Biol.*, 371 :1099–1117, 2007.

[72] O.V. Galzitskaya and A.V. Finkelstein. A theoretical search for folding/unfolding nuclei in three-dimensional protein sructures. *Proc. Natl. Acad. Sci. USA*, 96 :11299–11304, 1999.

[73] J. Gao, K. Kuczera, B. Tidor, and M. Karplus. Hidden thermodynamics of mutant proteins : A molecular dynamics analysis. *Science*, 244 :1069–1072, 1989.

[74] K. Gibson and H. A. Scheraga. Minimization of polypeptide energy. i. preliminary structures of bovine pancreatic ribonuclease s-peptide. *Proc. Natl. Acad. Sci. USA*, 58 :420, 1967.

[75] K. Gibson and H. A. Scheraga. Minimization of polypeptide energy. ii. preliminary structures of oxytocin, vasopressin, and an octapeptide from ribonuclease. *Proc. Natl. Acad. Sci. USA*, 58 :420, 1967.

[76] W. Gilbert. The rna world. *Nature*, 319 :618, 1986.

[77] N. Go. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.*, 12 :183–210, 1983.

[78] A. Godzik and J. Skolnick. Flexible algorithm for direct multiple alignement of protein structures and sequences. *Comput. Appl. Biosci.*, 6 :587–596, 1994.

[79] H. Gohlke, C. Kiel, and D.A. Case. Insight into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J. Mol. Biol.*, 330 :891–913, 2003.

[80] R.F. Goldstein. Efficient rotamer elimination applied to protein sidechains and related spin glasses. *Biophys. J.*, 66 :1335–1340, 1994.

[81] D.B. Gordon, S.A. Marshall, and S.L. Mayo. Energy functions for protein design. *Curr. Opin. Struct. Biol.*, 9 :509–513, 1999.

[82] R. Guerois and L. Serrano. Protein design based on folding models. *Curr. Opin. Struct. Biol.*, 11 :101–106, 2001.

[83] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The rna moiety of ribonuclease p is the catalytic subunit of the enzyme. *Cell*, 35 :849–857, 1983.

[84] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, 1998 :1462–1467, 1998.

[85] J.J. Havranek and P.B. Harbury. Tanford-Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. USA*, 96 :11145–11150, 1999.

[86] G.D. Hawkins, C. Cramer, and D. Truhlar. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem*, 100 :19824–19839, 1996.

[87] H. Hellinga and F. Richards. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. USA*, 91 :5803–5807, 1994.

[88] Z. Hendsch and B. Tidor. Electrostatic interactions in the GCN4 leucine zipper : substantial contributions arise from intramolecular interactions enhanced on binding. *Protein Science*, 8 :1381–1392, 1999.

[89] C.P. Hill, D.H. Anderson, L. Wesson, W.F. DeGrado, and D. Eisenberg. Crystal structure of alpha1 : Implications for protein design. *Science*, 249 :543–546, 1990.

[90] N. Hino, Y. Okazaki, T. Kobayashi, A. Hayashi, K. Sakamoto, and S. Yokoyama. Protein photo-cross-linking in mammalian cells by site-specific incorporation of a photoreactive amino acid. *Nat. Methods*, 2 :201–206, 2005.

[91] J.H. Holland. Adapatation in natural and artificial systems. *Cambridge, MA : The MIT Press*, - :–, 1992.

[92] L. Holland and C. Sander. Fast and simple monte carlo algorithm for side chain optimization in proteins : application to model building by homology. *Proteins*, 2 :213–223, 1992.

[93] Barry Honig and Anthony Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268 :1144–1149, 1995.

[94] C. Hountondji, P. Dessen, and S. Blanquet. Sequence similarities among the family of aminoacyl-trna synthetases. *Biochimie*, 68 :1071–1094, 1986.

[95] M. Illangasekare, G. Sanchez, T. Nickles, and M. Yarus. Aminoacyl-rna synthesis catalyzed by an rna. *Science*, 267 :643–647, 1995.

[96] H. Jakubowski and A.R. Fersht. Alternative pathways for editing non-cognate amino acids by aminoacyl-trna synthetases. *Nucleic Acids Research*, 9 :3105–3117, 1978.

[97] J. Janin, S. Wodak, M. Levitt, and B. Maigret. Conformation of amino acid side-chains in proteins. *J. Mol. Biol.*, 125 :357–386, 1978.

[98] A. Jaramillo, L. Wernisch, S. Héry, and S. Wodak. Folding free energy function selects native-like protein sequences in the core but not on the surface. *Proc. Natl. Acad. Sci. USA*, 99 :13554–13559, 2002.

[99] A. Jaramillo and S. Wodak. Computational protein design is a challenge for implicit solvation models. *Biophys. J.*, 88 :156–171, 2005.

[100] X. Jiang, H. Farid, E. Pistor, and R.S. Farid. A new approach to the design of uniquely folded thermally stable. *Protein Sci.*, 9 :403–416, 2000.

[101] D.T. Jones. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.*, 4 :567–574, 1994.

[102] W. Jorgensen and J. Tirado-Rives. The OPLS potential function for proteins, energy minimization for crystals of cyclic peptides and crambin . *J. Am. Chem. Soc.*, 110 :1657–1666, 1988.

[103] M. Karplus and A. Sali. Theoretical studies of protein folding and unfolding. *Curr. Opin. Struct. Biol.*, 5 :58–73, 1995.

[104] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Prot. Chem.*, 14 :1–63, 1959.

[105] T. Kawabata. Matras : a program for protein 3d structure comparison. *Nucleic Acids Research*, 13 :3367–3369, 2003.

[106] D.E. Kim, C. Fisher, and D. Baker. A breakdown of symmetry in the folding transition state of protein l. *J. Mol. Biol.*, 298 :971–984, 2000.

[107] J. Kirkwood and F. Westheimer. The electrostatic influence of substituents on the dissociation constant of organic acids. *J. Chem. Phys.*, 6 :506–512, 1938.

[108] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein sidechain conformations and estimate their conformational entropy. *J. Mol. Biol.*, 239 :249–275, 1994.

[109] P. Koehl and M. Delarue. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.*, 6 :222–226, 1996.

[110] P. Koehl and M. Levitt. De novo protein design. II. Plasticity in sequence space. *J. Mol. Biol.*, 293 :1183–1193, 1999.

[111] P.A. Kollman. Free energy calculations : applications to chemical and biochemical phenomena. *Chem. Rev.*, 93 :2395, 1993.

[112] H. Kono and J. Doi. Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins*, 19 :244–255, 1994.

[113] H. Kono and J.G. Saven. Statistical theory for protein combinatorial libraries. packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J. Mol. Biol.*, 306 :607–628, 2001.

[114] T. Kortemme, A. Morozov, and D. Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.*, 326 :1239–1259, 2003.

[115] C.M. Kraemer-Pecore, J.T. Lecomte, and J.R. Desjarlais. de novo redesign of the ww domain. *PS*, 12 :2194–2105, 2003.

[116] C.M. Kraemer-Pecore, A.M. Wollacott, and J.R. Desjarlais. Computational protein design. *Curr. Opin. Chem. Bio.*, 5 :690–695, 2001.

[117] K. Kruger, P.J. Grabowski, A.J. Zaug, J. Sands, D.E. Gottschling, and T.R. Cech. Self-splicing rna : autoexcision and autocyclization of the ribosomal rna intervening sequence of tetrahymena. *Cell*, 31 :147–157, 1982.

[118] B. Kuhlman and D. Baker. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA*, 97 :10383–10388, 2000.

[119] B. Kuhlman and D. Baker. Exploring folding free energy landscapes using computational protein design. *Curr. Opin. Struct. Biol.*, 14 :89–95, 2004.

[120] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302 :1364–1368, 2003.

[121] B. Kuhlman, JW. O'Neill, DE. Kim, KY. Zhang, and D. Baker. Accurate computer-based design of a new backbone conformation in the second turn of protein l. *J. Mol. Biol.*, 315 :471–477, 2002.

[122] V. Lamour, S. Quevillon, S. Diriong, V.C. N'Guyen, M. Lipinski, and M. Mirande. Evolution of the glx-trna synthetase family : the glutaminyl enzyme as a case of horizontal gene transfer. *Proc. Natl. Acad. Sci. USA*, 91 :8670–8674, 1994.

[123] J. Lapointe and R. Giege. *Transfer RNAs and aminoacyl-tRNA synthetases. Translation in Eukaryotes*. In trachsel, H. (ed.). CRC press Inc. Boca Raton, 1991.

[124] S.M. Larson, J. E. England, J.R. Desjarlais, and V.S. Pande. Thoroughly sampling sequence space : Large-scale protein design of structural ensembles. *Protein Science*, 11 :2804–2813, 2002.

[125] I. Lasters, M. De Maeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein sidechains. *Protein Eng.*, 8 :893–904, 1995.

[126] G.A. Lazar, J.R. Desjarlais, and T.M. Handel. De novo design of the hydrophobic core of ubiquitin. *Protein Science*, 6 :1167–1178, 1997.

[127] G.A. Lazar, S.A. Marsall, J.J. Plecs, S.L. Mayo, and J.R. Desjarlais. Designing proteins for therapeutic applications. *Curr. Opin. Struct. Biol.*, 13 :513–518, 2003.

[128] T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins*, 35 :133–152, 1999.

[129] R.J. Leatherbarrow, A.R. Fersht, and G. Winter. Transition-state stabilization in the mechanism of tyrosyl-trna synthetase revealed by protein engineering. *Proc. Natl. Acad. Sci. USA*, 82 :7840–7844, 1985.

[130] B. Lee and F. Richards. The interpretation of protein structures : estimation of static accessibility. *J. Mol. Biol.*, 55 :379–400, 1971.

[131] M.S. Lee, F.R. Salsbury Jr., and C.L. Brooks III. Constant pH molecular dynamics using continuous titration coordinates. *Proteins*, 56 :738–752, 2004.

[132] N. Lee, Y. Bessho, K. Wei, J.W. Szostak, and H. Suga. Ribozyme-catalyzed trna aminoacylation. *Nat. Struct. Biol.*, 7 :28–33, 2000.

[133] L.L. Looger and H.W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein sidechain structure prediction tractable : Implications for protein design and structural genomics. *J. Mol. Biol.*, 307 :429–445, 2001.

[134] A. Lopes, A. Aleksandrov, C. Bathelt, G. Archontis, and T. Simonson. Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins*, 67 :853–867, 2007.

[135] B.Z. Lu, W.Z. Chen, C.X. Wang, and X.J. Xu. Protein molecular dynamics with electrostatic force entirely determined by a single poisson-boltzmann calculation. *Proteins*, 48 :497–504, 2002.

[136] K.J. Lumb and P.S. Kim. A buried interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry*, 34 :8642–8648, 1995.

[137] M. De Maeyer, J. Desmet, and I. Lasters. All in one : a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.*, 2 :53–66, 1997.

[138] N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt, and A. Caflisch. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins*, 37 :88–105, 1999.

[139] S.M. Malakauskas and S.L. Mayo. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.*, 5 :470–475, 1998.

[140] S.A. Marshall and S.L. Mayo. Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.*, 305 :619–631, 2001.

[141] S.A. Marshall, C.S. Morgan, and S.L. Mayo. Electrostatic significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.*, 316 :189–199, 2002.

[142] S.A. Marshall, C.L. Vizcarra, and S.L. Mayo. One- and two-body decomposable poisson-boltzmann methods for protein design calculations. *Protein Science*, 14 :1293–1304, 2005.

[143] J.S. Marvin and H.W. Hellinga. Conversion of a maltose receptor into a zinc biosensor by computational design. *Proc. Natl. Acad. Sci. USA*, 98 :4955–4960, 2001.

[144] E.L. McCallister, E. Alm, and D. Baker. Critical role of β-hairpin formation in protein g folding. *Nat. Struct. Biol.*, 7 :669–673, 2000.

[145] J.A. McCammon, B. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267 :585, 1977.

[146] J.A. McCammon and S. Harvey. *Dynamics of proteins and nucleic acids*. Cambridge University Press, Cambridge, 1987.

[147] R.A. Mehl, J.C. Anderson, S.W Santoro, L. Wang, A.B. Martin, D.S. Kings, D.M. Horn, and P.G. Schultz. Generation of a bacterium with a 21 amino acid genetic code. *J. Am. Chem. Soc.*, 125 :935–939, 2003.

[148] J. Mendes, A.M. Baptista, M.A. Carrondo, and C.M. Soares. Improved modeling of side-chains in proteins with rotamer-based methods : a flexible rotamer model. *Proteins*, 37 :30–543, 1999.

[149] J. Mendes, A.M. Baptista, M.A. Carrondo, and C.M. Soares. Implicit solvation in the self-consistent mean field theory method : sidechain modelling and prediction of folding free energy of protein mutants. *J. Comput. Aided Mol. Des*, 15 :721–740, 2002.

[150] J. Mendes, R. Guerois, and L. Serrano. Energy estimation in protein design. *Curr. Opin. Struct. Biol.*, 12 :441–446, 2002.

[151] J.S. Merkel and L. Regan. Modulating protein folding rates in vivo and in vitro by side-chain interactions between the parallel βstrands of green fluorescent protein. *J. Biol. Chem.*, 275 :29200–29206, 2000.

[152] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, and A.H. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21 :1087–1092, 1953.

[153] B. H. M. Mooers, D. Datta, W. A. Baase, E. S. Zollars, S. L. Mayo, and B. W. Matthews. Repacking the core of t4 lysozyme by automated design. *J. Mol. Biol.*, 6 :741–756, 2003.

[154] L. Moulinier, S. Eiler, G. Eriani, J. Gangloff, J.C. Thierry, K. Gabriel, W.H. McClain, and D. Moras. The structure of an AspRS-tRNA(asp) complex reveals a tRNA-dependent control mechanism. *EMBO J.*, 20 :5290, 2001.

[155] V. Munoz and W.A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA*, 96 :11311–11316, 1999.

[156] M. Munson, K.S. Anderson, and L. Regan. Speeding up protein folding : mutations that increase the rate at which rop folds and unfolds. *Fold. Des*, 2 :77–87, 1997.

[157] A.G. Murzin, A.M. Lesk, and C. Chothia. Principles determining the structure of beta-sheet barrels in proteins. i. a theoretical analysis. *J. Mol. Biol.*, 236 :1369–1381, 1994.

[158] A.G. Murzin, A.M. Lesk, and C. Chothia. Principles determining the structure of beta-sheet barrels in proteins. ii. the observed structures. *J. Mol. Biol.*, 236 :1382–1400, 1994.

[159] G.M. Nagel and R.F. Doolittle. Phylogenetic analysis of the aminoacyl-trna synthetases. *J. Mol. Evol.*, 40 :487–498, 1995.

[160] T. Nakatsu, H. Kato, and J. Oda. Crystallization and preliminary crystallographic study of asparagine synthetase from escherichia coli. *Acta. Crystallogr. D. Biol. Crystallogr.*, 52 :604–606, 1996.

[161] T. Nakatsu, H. Kato, and J. Oda. Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class ii aminoacyl-trna synthetase. *Nat. Struct. Biol.*, 5 :15–9, 1998.

[162] S. Ohnishi, A.L. Lee, M.H. Edgell, and D. Shortle. Direct demonstration of structural similarity between native and denaturated eglin c. *Biochemistry*, 43 :4064–4070, 2004.

[163] A. Onufriev, D. Bashford, and D.A. Case. Modification of the generalized Born model suitable for macromolecules. *J. Phys. Chem. B*, 104 :3712–3720, 2000.

[164] T. Ooi, M. Oobatake, G. Nemethy, and H. Scheraga. Accessible surface areas as a measure of the thermodynamic hydration parameters of peptides. *Proc. Natl. Acad. Sci. USA*, 84 :3086–3090, 1987.

[165] N. Pokala and T. M. Handel. Energy functions for protein design : Adjustement with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.*, 347 :203–227, 2005.

[166] N. Pokala and T.M. Handel. Protein design : Where we were, where we are, where we're going. *Journal of Structural Biology*, 134 :269–281, 2001.

[167] N. Pokala and T.M. Handel. Energy functions for protein design I : Efficient and accurate continuum electrostatics and solvation. *Protein Science*, 13 :925–936, 2004.

[168] J. Ponder and F.M. Richards. Tertiary templates for proteins : Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193 :775–791, 1987.

[169] A. Poterszman, M. Delarue, J.C. Thierry, and D. Moras. Synthesis and recognition of aspartyl-adenylate by *Thermus thermophilus* aspartyl-tRNA synthetase. *J. Mol. Biol.*, 244 :158–167, 1994.

[170] K. Raha, A. M. Wollacott, M. J. Italia, and J. R. Desjarlais. Prediction of amino acid sequence from structure. *Protein Science*, 9 :1106–1119, 2000.

[171] M. Ramirez-Alvarado, J.S. Merkel, and L. Regan. A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. *Proc. Natl. Acad. Sci. USA*, 97 :8979–8984, 2000.

[172] C.S. Rapp and R.A. Friesner. Prediction of loop geometry using a Generalized Born model of solvation effects. *Proteins*, 35 :173–183, 1999.

[173] L. Ribas de Pouplana and P. Schimmel. A view into the origin of life : aminoacyl-trna synthetases. *Cell. Mol. Life Sci.*, 57 :865–870, 2000.

[174] L. Ribas de Pouplana and P. Schimmel. Aminoacyl-trna synthetases : potential markers of genetic code development. *Trends. Biochem. Sci.*, 26 :591–596, 2001.

[175] L. Ribas de Pouplana and P. Schimmel. Two classes of trna synthetases suggested by sterically compatible dockings on trna acceptor stem. *Cell*, 104 :191–193, 2001.

[176] M.G. Rossmann, D. Moras, and K.W. Olsen. Chemical and biological evolution of nucleotide-binding protein. *Nature*, 250 :194–199, 1974.

[177] M. Rould, J. Perona, D. Söll, and T.A. Steitz. Structure of *Escherichia coli* glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 å resolution. *Science*, 246 :1135–1142, 1989.

[178] B. Roux. Theoretical and comptational models of ion channels. *Curr. Opin. Struct. Biol.*, 12 :182–189, 2002.

[179] K. Sakamoto, Hayashi A, Sakamoto A, Kiga D, Nakayama H, Soma A, Kobayashi T, Kitabatake M, Takio K, Saito K, Shirouzu M, Hirao I, and Yokoyama S. Site-specific incorporation of an unnatural amino acid into proteins in mammalian cells. *Nucleic Acids Research*, 30 :4692–4699, 2002.

[180] C.T. Saunders and D. Baker. Recapitulation of protein family divergence using flexible backbone protein design. *J. Mol. Biol.*, 346 :631–644, 2005.

[181] M. Schaefer and M. Karplus. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem*, 100 :1578–1599, 1996.

[182] P. Schimmel, R. Giege, D. Moras, and S. Yokoyama. An operational rna code for amino acids and possible relationship to genetic code. *Proc. Natl. Acad. Sci. USA*, 90 :8763–8768, 1993.

[183] P. Schimmel and S.O. Kelley. Exiting an rna world. *Nat. Struct. Biol.*, 7 :5–7, 2000.

[184] P. Schimmel and D. Soll. Aminoacyl-trna synthetases : general features and recognition of transfer rnas. *Annu Rev Biochem*, 48 :601–648, 1979.

[185] M. Schmidt am Busch, A. Lopes, D. Mignon, and T. Simonson. Computational protein design : software implementation, parameter optimization, and performance of a simple model. *J. Comput. Chem.*, 29 :1092–102, 2008.

[186] M. Schmidt am Bush, A. Lopes, N. Amara, C. Bathelt, and T. Simonson. Testing the coulomb/accessible surface area solvent model for protein stability, ligand binding, and protein design. *BMC Bioinformatics*, 9 :148, 2008.

[187] E. Schmitt, T. Meinnel, S. Blanquet, and Y. Mechulam. Methionyl-trna synthetase needs an intact and mobile 332kmsks336 motif in catalysis of methionyl adenylate formation. *J. Mol. Biol.*, 242 :566–576, 1994.

[188] E. Schmitt, L. Moulinier, S. Fujiwara, T. Imanaka, J.C.Thierry, and D. Moras. Crystal structure of aspartyl-trna synthetase from pyrococcus kodakaraensis kod : archaeon specificity and catalytic mechanism of adenylate formation. *EMBO J.*, 17 :5227–5237, 1998.

[189] B.K. Shoichet, W.A. Baase, R. Kuroki, and B.W. Matthews. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. USA*, 92 :452–456, 1995.

[190] T. Simonson, J. Carlsson, and D. A. Case. Proton binding to proteins : pK_a calculations with explicit and implicit solvent models. *J. Am. Chem. Soc.*, 126 :4167–4180, 2004.

[191] T. Simonson, D. Mignon, M. Schmidt am Busch, A. Lopes, and C. Bathelt. *The inverse protein folding problem : structure prediction in the genomic era.* Tektum Publishers, Berlin, 2007.

[192] A.G. Street, D. Datta, D.B. Gordon, and S.L Mayo. Designing protein β-sheet surfaces by z-score optimization. *Physical Review Letters*, 84 :5010–5013, 2000.

[193] A.G. Street and S.L Mayo. Pairwise calculation of protein solvent-accessible surface areas. *Folding and Design*, 3 :253–258, 1998.

[194] A. Stromgaard, A.A. Jensen, and K. Stromgaard. Site-specific incorporation of unnatural amino acids into proteins. *ChemBioChem*, 5 :909–916, 2004.

[195] A. Su and S.L. Mayo. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Science*, 9 :1701–1707, 1997.

[196] N. Taddei, F. Chiti, T. Fiaschi, M. Bucciantini, C. Capanni, M. Stefani, L. Serrano, C.M. Dobson, and G. Ramponi. Stabilisation of alpha-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphophatase. *J. Mol. Biol.*, 300 :633–647, 2000.

[197] C. Tanford and J. Kirkwood. Theory of protein titration curves. General equations for impenetrable spheres. *J. Am. Chem. Soc.*, 79 :5333–5339, 1957.

[198] D. Thompson, C. Lazennec, P. Plateau, and T. Simonson. Ammonium scanning in an enzyme active site : the chiral specificity of aspartyl-tRNA synthetase. *The Journal of Biological Chemistry*, in press :0000, 2007.

[199] D. Thompson, P. Plateau, and T. Simonson. Free energy simulations reveal long-range electrostatic interactions and substrate-assisted specificity in an aminoacyl-tRNA synthetase. *ChemBioChem*, 7 :337–344, 2006.

[200] D. Thompson and T. Simonson. Molecular dynamics simulations show that bound mg^{2+} contributes to amino acid and aminoacyl adenylate binding specificity in aspartyl-trna synthetase through long range electrostatic interactions. *The Journal of Biological Chemistry*, 281 :23792–23803, 2006.

[201] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *Journal of Biomolecular Structure & Dynamics*, 8 :1267–1289, 1991.

[202] S. Ventura, Vega M. C., Lacroix E., Angrand I., Spagnolo L., and Serrano L. Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.*, 9 :485–493, 2002.

[203] A.R. Viguera, V. Villegas, F.X. Aviles, and L. Serrano. Favourable native-like helical local interactions can accelerate protein folding. *Fold. Des.*, 2 :23–33, 1997.

[204] V. Villegas, J. Zurdo, V.V. Filimonov, F.X. Aviles, C.M. Dobson, and L. Serrano. Protein engineering as a strategy to avoid formation of amyloid fibrils. *Protein Science*, 9 :1700–1708, 2000.

[205] C.L. Vizcarra and S.L Mayo. Electrostatic in computational protein design. *Curr. Opin. Struct. Biol.*, 9 :622–626, 2005.

[206] C.A. Voigt, D.B. Gordon, and S.L Mayo. Trading accuracy for speed : A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.*, 299 :789–803, 2000.

[207] L. Wang, A. Brock, B. Herberich, and P.G. Schultz. Expanding the genetic code of escherichia coli. *Science*, 292 :498–500, 2001.

[208] L. Wang and P.G. Schultz. Expanding the genetic code. *Angew. Chem. Int. Ed. Engl.*, 44 :34–66, 2004.

[209] L. Wang, Z. Zhang, A. Brock, and P.G. Schultz. Addition of the keto functional group to the genetic code of escherichia coli. *Proc. Natl. Acad. Sci. USA*, 100 :56–61, 2001.

[210] T.A. Webster, H. Tsai, M. Kula, G.A. Mackie, and P. Schimmel. Specific sequence homology and three dimensional structure of an aminoacyl transfer rna synthetase. *Science*, 226 :1315–1317, 1984.

[211] L. Wernisch, S. Hery, and S. Wodak. Automatic protein design with all atom force fields by exact and heuristic optimization. *J. Mol. Biol.*, 301 :713–736, 2000.

[212] L. Wesson and D. Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, 1 :227–235, 1992.

[213] C.R. Woese, G.J. Olsen, M. Ibba, and D. Soll. Aminoacyl-trna synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.*, 64 :202–236, 2000.

[214] J. Xie and P.G. Schultz. Adding amino acids to the genetic repertoire. *Curr. Opin. Chem. Biol.*, 9 :548–554, 2005.

[215] J.M. Yang, C.H. Tsai, M.J. Hwang, H.K. Tsai, J.K. Hwang, and C.Y. Kao. GEM : A gaussian evolutionary method for predicting protein sidechain conformations. *Protein Science*, 11 :1897–1907, 2002.

[216] ES. Zollars, SA. Marshall, and SL. Mayo. Simple electrostatic model improves designed protein sequences. *Protein Science*, 15 :2014–2018, 2006.

Liste des publications

[1] N. Papandreou, I. N. Berezovsky**, A. Lopes**, E. Eliopoulos and J. Chomilier, 2004. Universal positions in globular proteins. From observation to simulation, *Europeen Journal of Biochemistry, 271, 4762-4768* .

[2] **A. Lopes**, A. Alexandrov, C. Bathelt, G. Archontis and T. Simonson, 2007. Computational sidechain placement and protein mutagenesis with implicit solvent models. *PROTEINS: Structure, Function and Bioinformatics, 67, 853-867.*

[3] M. S. am Busch, **A. Lopes,** D. Mignon, and T. Simonson, 2007. Computational protein design: software implementation, parameter optimization, and performance of a simple model**,** 2008**.** *Journal of Computational Chemistry, 29, 1092-102.*

[4] M. S. am Busch**, A. Lopes**, C. Bathelt, N. Amara and T. Simonson, 2008. Testing the Coulomb/Accessible Surface Area solvent model for protein stability, ligand binding, and protein design, 2008. *BMC Bioinformatics, 9:148.*

[5] T. Simonson, D. Mignon, M. S. am Busch, **A. Lopes**, C. Bathelt, 2007. The inverse protein folding problem: structure prediction in the genomic era. Distributed & Grid Computing - Science Made Transparent for Everyone. Principles, Applications and Supporting Communities, *Tektum Publishers, Berlin. (Book chapter)*

[6] **A. Lopes**, M. S. am Busch and T. Simonson, 2008. (in preparation)

Chapitre 8

Remerciements

Tout d'abord je tiens à remercier chacun des membres du jury, Jacques Chomilier, Marc Delarue, Gilbert Deléage, Raphaël Guerois, Martin Karplus et Pierre Tufféry pour avoir accepté de juger mon travail. Je remercie aussi Jacques Chomilier pour l'encadrement de qualité qu'il m'a offert lors de mon stage de DEA et pour son soutien et nombreux conseils. Je tiens particulièrement à remercier Thomas Simonson, tout d'abord pour la qualité de l'encadrement dont j'ai pu bénéficier sous sa direction. En particulier, j'ai grandement profité de sa rigueur scientifique et de l'ampleur du projet qu'il m'a confié. Justement, je le remercie aussi de la confiance qu'il m'a témoignée tout au long de ma thèse en me donnant un sujet très riche et me laissant organiser mon travail comme je l'entendais. Enfin, je le remercie pour son précieux soutien et son accompagnement tout au long de mes démarches de recherches de post-doc. Merci aussi à Giorgios Archontis pour le travail que nous avons accompli ensemble et la sérénité qu'il diffuse autour de lui. Je remercie également Yves Mechulam pour m'avoir confié un poste de moniteur sur l'école ; cette expérience fût très enrichissante pour moi. Je remercie également Pierre Plateau pour m'avoir accueillie dans son laboratoire et pour les bons souvenirs que je garde des oraux que nous avons fait passer ensemble aux élèves.

Je remercie aussi Franca Fraternali pour m'accorder sa confiance et se démener dans les recherches de financements. Comme tu le dis si souvent 'Let's try. Let's hope' !

Mes remerciements vont ensuite pour l'ensemble des personnes du laboratoire qui ont contribué à une superbe ambiance pendant ces trois années : Alexey, Alfonso, Annick,

Carotte, Catherine (je repasserai te voir quand j'aurais besoin d'enveloppes pour payer mes amendes ! ! !), les deux Christine, David, Emmanuelle, Laura, Laurent G., Lionel, Lolo, Marc, Marcel, Marie, Maria, Mélanie, Michel, Michou, Myriam, Najette, Pablo, Pascal, Sandra (je repasserai vous faire peur à toi et Caro quand vous travaillerez tard dans la nuit !) et Sylvain. Une pensée particulière pour mes deux compères bioinformaticiens Guillaume L. et Josselin ! Merci Guillaume pour m'avoir aidé à garder la ligne en mangeant tous mes bonbons et plus sérieusement pour tes nombreux conseils de 'grand frère' tout au long de ma thèse. 'Big thank's' for Joss ! J'ai passé le mot sur Ivry, tu peux y laisser ton scooter, il ne lui arrivera rien ! ! ! Je n'oublie pas Laurent M. qui m'a accompagné de nombreuses fois dans mes repas tardifs lorsque je restais tard au labo ! J'en profite pour remercier Bertrand de 'Phyleas' pour tous les brownies et chocolats qu'il m'a offert pour m'encourager pendant la rédaction ! Enfin un grand MERCI à Guillaume H. et Romary, mes amis avec qui nous avons passé d'excellentes soirées bien tardives 'et même que c'est pas fini' ! ! !

Un grand MERCI à ma famille du DEA avec qui j'ai partagé le stress et les longues heures de travail lors des rédactions de projets, avec qui j'ai traversé le froid (au ski !), avec qui j'ai subi la cantine de Jussieu, avec qui j'ai englouti les lasagnes maisons (ça c'était meilleur !) et avec qui je continue de partager de superbes moments et fous rires ! Donc un grand MERCI à Adrien parti chez les 'oranges', Emmanuel (honte sur nous, en trois ans nous ne sommes même pas venus te voir à Cambridge), 'Futur' (prépare toi, on débarque bientôt te voir en Sardaigne, futur pays de la recherche !), Gaëlle (à quand la prochaine 'session ski' ?), Guillaume, Hocine ou Dr MAD. (on attend toujours les religieuses ! En tout cas merci pour les nombreuses rotondes du CEA !), Julie et Valérie (ce sera quelque chose si on se retrouve toutes les trois en colloque : on fait un deal, vous payez le loyer, je fais la cuisine ! Je ne sais pas qui y gagne ! ! !), Julien et Stéphanie (promis je cherche encore ton stylo bille 6 couleurs !), Karine (pour tous les fous rires : j'en ai encore mal au ventre !), Marie, Maxime (merci pour ta générosité, toujours là, prêt à donner un coup de main ! Présent pour tous les plans, merci pour tous ces souvenirs !), Nours (je te dois encore une boule de neige), Thien-An et Thomas. Merci aussi aux enseignants d'AGM2 à qui je dois une formation très solide. J'ai énormément appris au cours de ce DEA, que

ce soit au niveau de la rigueur scientifique, de l'ouverture d'esprit ou de l'autonomie de travail. Une pensée particulière pour Delphine Flatters et Patrick Fuchs. Merci pour votre soutien et vos précieux conseils...

Un grand MERCI à ma famille d'Ivry : mes amis d'enfance, toujours les mêmes, toujours unis. Merci à Christine et Hanafi (pour tous ces superbes repas ! Rendez vous un de ces jours en Espagne ou Algérie pour changer d'Ivry !), Fazia (éternelle amie d'enfance, toujours présente, merci pour ton soutien, ton éternelle joie et ton sourire), Julie (je triche un peu, tu nous as abandonné pour aller te réchauffer à Marseille !), Karine (ma seconde grande soeur, merci pour tes conseils et comme promis rendez vous chez 'jénial' !), Lydia (comme pour Faz, toujours là depuis le début, merci de m'avoir apporté du couscous au labo quand je passais les nuits là bas ! C'est vrai que c'était meilleur quand j'ai trouvé où se trouvait le micro-onde !). Enfin, je remercie particulièrement Mr Castille, professeur de physique au lycée Romain Rolland, pour m'avoir appris la rigueur et fait aimer la chimie.

Je n'oublie pas la famille Ndong/Ndour (Sophie, Quand tu veux pour ton Mafé !), Chainez (toujours chargée de cadeaux et de bons desserts !), Jean Hubert (merci pour tes cours d'anglais, j'en ai bien besoin !), Arnaud Cotel (pour les petites pauses café !) et Arnaud Méru pour tes nombreux conseils...

Une pensée pour ma famille d'Espagne : José et Pacucha, Paquita, une pensée singulière pour Juan Carlos, Miriam, José Manuel et le magnifique petit Ruben ; j'aurais vraiment aimé vous voir lors de ma soutenance, je vous la referais en espagnol l'été prochain ! Enfin, une pensée spéciale pour mes deux 'petites soeurs' espagnoles : Yasmina et Romina (promis je t'emmène à Paris l'an prochain !).

Je n'oublie pas Sergio et Virginie, ma cousine toujours présente, je suis trop fière de toi, tout le monde n'a pas une cousine qui faisait partie des 'Berruriers Noirs' ! De même, j'ai une pensée pour mon petit cousin Alex (même si tu me dépasses d'une tête maintenant !).

Enfin, un grand MERCI, à ma famille qui m'a toujours soutenu dans tous mes choix. Merci à mes parents pour leurs conseils, pour m'avoir appris la valeur du travail, pour m'avoir poussé à toujours aller jusqu'au bout et poussé à me battre pour réaliser mes rêves. Merci à ma soeur (la vraie cette fois !) et à Laurent. Merci pour vos conseils,

vos valeurs et merci pour le mixeur !!! Une pensée pour mes deux neveux que j'adore plus que tout, un gros bisou pour Karlito et Pavel de la part de Ban-Anne ! Merci à la famille Cleuziou qui s'est toujours montrée très attentive. Enfin un grand MERCI à mon p'tit Thieum, Matthieu qui m'a toujours soutenu depuis le début et surtout jusqu'à la fin ! La fin s'est montrée bien difficile, merci d'avoir été présent lors de mes nombreuses inquiétudes et encore une fois merci pour les innombrables corrections de mon anglais lors de mes rédactions de projets de financement !

Merci à tous...

ABSTRACT

Protein design aims to develop new proteins with new structural and/or functional properties. The principle is to identify among the available sequences, those that preserve the protein's three dimensional fold and that confer the desired properties. The general procedure can be decomposed into two steps : (i) the interaction energy between all the residue pairs is precomputed and stored in an energy matrix taking into account all amino acid types and all possible conformations, (ii) an optimization algorithm explores simultaneously sequence and conformational space to determine the best amino acid combination. Next, different filters (based on affinity, protein stability…) can be applied to separate functional sequences (for a given fold) from the non functional ones.

We first focused on the development of the protein design procedure, particularly, on the setting up and optimization of the energy function and the implementation of the optimization algorithm. We have shown that our procedure is robust because it performs well for a wide variety of applications crucial in protein design such as: prediction of sidechain orientation, prediction of stability or affinity changes due to point mutations and design of native-like sequences for a set of globular proteins. For all of these applications the quality of results is competitive with those obtained by other groups.

Next, we applied our procedure to more complex systems such as protein:ligand complexes. We focused on aspartyl-tRNA synthetase (AspRS) and asparaginyl-tRNA synthetase (AsnRS). These enzymes play a crucial role in preserving the accuracy of genetic code translation, linking their specific amino acid to a cognate tRNA, which carries the corresponding anticodon. First, we performed the design of the whole active site of AspRS and AsnRS in presence of their specific or non specific ligands in order to test the performance of our procedure. The quality of the designed sequences is consistent with those observed on entire globular proteins. On the other hand, we showed that our procedure was sensitive to the nature of the ligand present in the active site. Finally, we performed the design of a limited number of selected positions of the AsnRS active site so that the synthetase can bind preferentially aspartate over asparagine. A set of promising mutants has been retained. Their stability and affinity for native and non-native ligands are now analyzed by molecular dynamics.