Académie de Montpellier

Université Montpellier II
— Sciences et Techniques du Languedoc —

Thèse

pour obtenir le grade de

Docteur de l'université Montpellier II

Discipline : **Informatique**
Ecole doctorale : **Information, Structures, Systèmes**

Optimisation de potentiels statistiques pour un modèle d'évolution soumis à des contraintes structurales

Présentée et soutenue publiquement par

Cécile Bonnard

le 5 janvier 2010

Jury

Asger Hobolth, associate professor, Aarhus University ..... Rapporteur
Thomas Simonson, professeur, Polytechnique ..... Rapporteur
Jérôme Gracy, ingénieur de recherche, CNRS ..... Examinateur
Yann Guermeur, chargé de recherche, CNRS ..... Examinateur
Olivier Gascuel, directeur de recherche, LIRMM ..... Directeur de thèse
Nicolas Lartillot, professeur adjoint, Université de Montréal ..... Co-directeur de thèse

Remerciements

Tout d'abord, je tiens à remercier mon co-directeur de thèse, Nicolas Lartillot, pour m'avoir permis de faire ma thèse avec lui, pour sa patience et son encadrement. Ce fut un plaisir et un honneur que d'avoir pu travailler avec lui, sur un sujet pas évident au départ, mais qui s'est révélé passionnant. Je remercie également mon autre co-directeur de thèse, Olivier Gascuel, pour sa générosité, son aide, et ses remarques sur mon document de thèse.

Je remercie Claudia L. Kleinman et Nicolas Rodrigue, avec qui j'ai eu la chance et le plaisir de travailler, et Hervé Philippe, pour m'avoir accueillie à plusieurs reprises dans son laboratoire.

Je remercie mes deux rapporteurs, Asger Hobolth et Thomas Simonson, ainsi que les membres de mon jury, Yann Guermeur et Jérôme Gracy. Ce dernier, ainsi que Alain Jean-Marie étaient membres de mon comité de thèse, et je tenais à les remercier pour avoir suivi mon travail pendant ces années, et pour les conseils qu'ils m'ont apporté. Je remercie également Gilles Caraux, notamment pour ses conseils qui m'ont permis de m'orienter vers la bioinformatique, et Yolande Ahronovitz pour m'avoir permise de donner des cours dans sa matière, et qui m'a donné le goût de l'enseignement.

Je souhaite remercier la Région Languedoc-Roussillon, le CNRS, l'université de Montpellier II et l'Université de Montréal, pour les différents financements qui m'ont permis d'effectuer ma thèse. Je tiens à remercier particulièrement le personnel de l'université de Montpellier II, et notamment Bernadette et Olivia, pour leur aide précieuse pour l'accomplissement de toutes les démarches administratives afin que je puisse soutenir dans les temps. Je souhaite également remercier Nicole, Cécile et Elisabeth, pour leur disponibilité, leur sourire et leur gentillesse. J'ai aussi une pensée pour Nadine Tilloy, qui nous a hélas quittés il y a peu de temps, et qui m'a également aidée dans mes démarches administratives.

Je remercie les doctorants de Montpellier : Raluca, Flop, Ben, Gilles, Céline, Denis, Olivier, Fred, JB, Nicolas, Sam, et j'en oublie... Mais également les membres de l'équipe MAB, Annie, Laurent, Séverine, Valentin, Eric, Alban, Vincent, Jean-François...

Je remercie toutes les personnes que j'ai eu la chance de rencontrer à Montréal, dans le laboratoire, notamment : Nan, Bats, Marie, Pierre, Shen, Do, Véronique... Il est très facile de s'habituer à un nouveau pays, quand on a la chance de pouvoir vous rencontrer !

Je remercie mes amis, dispersés aux quatre coins du monde, pour leur soutien et leur amitié. Pour les amis les plus récents comme les plus anciens, ceux que je ne vois hélàs que peu mais qui restent près du coeur, comme ceux que j'ai la chance de voir souvent. Merci pour les voyages, les vacances, les jeux, les week-end, et les fous-rires, pour les dégustations de scotch et pour les soirées passées à s'amuser et rire sur skype. Notamment - car je n'aurais jamais assez de temps ni de place pour tous les remercier comme il se doit - merci à Suzanne, Aurélie et Marine, pour les voyages, Caro et Gaëlle pour leur compréhension, Jean pour son inébranlable gentillesse, et à Sylvie pour m'avoir fait l'honneur d'être son témoin. Enfin, je remercie particulièrement Mika pour avoir supporté stoïquement mes différentes sautes d'humeur, tout en continuant à m'encourager, et Lionel pour son soutien indéfectible.

Je remercie également tous ceux de ma famille qui m'ont soutenue, avec une pensée spéciale pour mes grand-pères décédés durant ma thèse, pour leur conseils judicieux, leur support et leur affection. J'aurais aimé pouvoir partager avec eux le plaisir et la fierté d'avoir terminé mon doctorat.

Mes derniers remerciements, mais non des moindres vont à mes parents, qui m'ont toujours aidée pendant ces (parfois) difficiles moments d'études. Cette thèse leur est dédiée, pour leur soutien, leur amour et la confiance sans faille qu'ils m'ont toujours apporté.

Table des matières

Préambule **7**

Introduction **11**

I Etat de l'art **19**

1 Modèles d'évolution probabilistes **21**

1.1 Introduction 21

1.2 Modèles de substitutions 23

1.2.1 Modèle de substitution markovien 23

1.2.2 Vraisemblance 25

1.2.3 Modèles d'évolution nucléotidiques 26

1.2.4 Modèles d'évolution à acides aminés 27

1.2.5 Modèles d'évolution à codons 31

1.3 Modèles mécanistiques 33

1.3.1 Modèle d'évolution mutation/sélection 33

1.3.2 Modèle d'évolution soumis à des contraintes structurales 35

1.3.3 Implémentation d'un modèle SC 39

2 Modéliser la sélection dépendant de la structure - Potentiels statistiques **41**

2.1 Introduction 41

2.1.1 Principes généraux 42

2.1.2 Espace des structures 43

2.2 Quelle forme d'énergie ? 47

2.2.1 Quelques propriétés 49
2.2.2 Champs de force semi-empiriques 50
2.2.3 Potentiels statistiques 51
2.2.4 Formes générales de potentiels statistiques 53
2.2.5 Estimer des énergies du potentiel 59
2.2.6 Optimisation directe des énergies 63
2.3 Le problème du *protein design* 70
2.3.1 Introduction 70
2.3.2 *Protein design* versus *protein folding* 71
2.3.3 Le *random energy model* 73
2.3.4 Optimisation versus échantillonnage 73
2.3.5 Optimisation de potentiels dans un contexte de *protein design* 74

3 Méthodes numériques et statistiques 77
3.1 Introduction et notations 77
3.2 Algorithme de Metropolis-Hasting 78
3.3 Algorithme d'échantillonnage de Gibbs 79
3.4 Méthode de descente de gradient 80
3.5 Facteur de Bayes 82
3.6 Conclusion 84

II Optimisation de potentiels pour le protein design et l'évolution moléculaire 87

4 Développement du cadre statistique 89
4.1 Introduction 89
4.2 Article 90
4.2.1 Abstract 90
4.2.2 Background 91
4.2.3 Results 94
4.2.4 Discussion 105
4.2.5 Conclusions 111
4.2.6 Methods 112

4.2.7 Authors' contributions . . . 116
4.2.8 Acknowledgements . . . 116
4.3 Conclusion . . . 116

5 Optimisation des potentiels à l'aide d'une pseudo-vraisemblance 119
5.1 Introduction . . . 119
5.2 Article . . . 120
5.2.1 Abstract . . . 120
5.2.2 Background . . . 121
5.2.3 Results . . . 123
5.2.4 Discussion . . . 133
5.2.5 Conclusions . . . 134
5.2.6 Methods . . . 135
5.2.7 Authors contributions . . . 139
5.2.8 Acknowledgements . . . 139
5.3 Conclusion . . . 139

6 Reformulation du probleme 143
6.1 Introduction . . . 143
6.2 Article . . . 144
6.2.1 Abstract . . . 144
6.2.2 Background . . . 145
6.2.3 Results . . . 147
6.2.4 Phylogenetic analysis . . . 153
6.2.5 Conclusions . . . 154
6.2.6 Methods . . . 155
6.3 Conclusion . . . 156

7 Inclusion de structures leurres 157
7.1 Introduction . . . 157
7.2 Article . . . 158
7.2.1 Abstract . . . 158
7.2.2 Background . . . 159
7.2.3 Methods . . . 161
7.2.4 Results . . . 171

7.2.5 Discussion 178
7.3 Conclusion 180

III Bilan de l'approche **183**

8 Perspectives **185**
8.1 Directions futures 185
8.1.1 Affiner le terme d'interaction 185
8.1.2 Amélioration de l'approche par structures leurres 188
8.1.3 Optimisation de potentiels statistiques dans un modèle d'évolution 192
8.2 Applications 194
8.3 Conclusion 197

Conclusions **199**

201

Annexes **203**

A Liste des abbréviations **203**

B Développement du cadre statistique **205**
B.1 Fichier additionnel 1 205
B.2 Fichier additionnel 2 206
B.3 Fichier additionnel 3 207
B.4 Fichier additionnel 4 209
B.5 Fichier additionnel 5 231
B.6 Fichier additionnel 6 233
B.7 Fichier additionnel 7 235
B.8 Fichier additionnel 8 250

C Optimisation des potentiels à l'aide d'une pseudo-vraisemblance **253**
C.1 Fichier additionnel 1 253
C.2 Fichier additionnel 2 254

C.3 Fichier additionnel 3 254
C.4 Fichier additionnel 4 256
C.5 Fichier additionnel 5 257
C.6 Fichier additionnel 6 257

Bibliographie **259**

Préambule

Lors de ma dernière année d'école d'ingénieur agronome, j'ai eu l'opportunité de réaliser un DEA d'informatique à l'université de Montpellier 2. Même s'il s'agissait d'un parcours apparemment atypique, par rapport à ma formation initiale, j'ai choisi ce DEA car je souhaitais m'orienter vers la bioinformatique, et cela me permettait d'avoir une certaine expertise dans le domaine informatique. Mon intérêt pour la bioinformatique m'a tout naturellement conduite à accomplir mon stage de DEA au sein de l'équipe Méthodes et Algorithmes pour la Biologie (MAB) au LIRMM, à Montpellier. Cette équipe s'intéresse à différents axes de recherche comme la création de méthodes informatiques pour l'alignement de séquences, la reconstruction phylogénétique, les modèles d'évolution... Effectué sous la direction de Nicolas Lartillot et Vincent Berry, mon stage portait sur les consensus multipolaires d'arbres phylogénetiques [Bonnard et al., 2006], et a contribué à me conforter dans ma décision de continuer dans le domaine de la bioinformatique.

Début novembre 2005, j'ai eu la chance de commencer une thèse sous la direction de Nicolas Lartillot et Olivier Gascuel, avec une bourse BDI (docteur ingénieur) co-financée par le CNRS et la région Languedoc-Roussillon. Le sujet de ma thèse porte sur l'optimisation de potentiels statistiques pour un modèle d'évolution soumis à des contraintes structurales. Ce modèle [Rodrigue et al., 2005,Rodrigue et al., 2006,Rodrigue et al., 2009] a été développé par Nicolas Rodrigue, à Montréal, et avec lequel Nicolas Lartillot collabore étroitement. Ce sujet m'intéressait tout particulièrement, car j'avais déjà eu la possibilité, en 2004, de faire un stage à Montréal, sous la direction de Nicolas Rodrigue et Hervé Philippe, et de participer alors à la naissance du projet.

De son côté, Claudia L. Kleinman commençait une thèse en reprenant les travaux que j'avais déjà effectués, ce qui donna lieu à un premier article [Kleinman et al., 2006] (présenté ici dans le chapitre 4) et à notre première collaboration. Par la suite, Nicolas Lartillot se vit offrir un poste à l'Université de Montréal, et ce fut pour moi l'opportunité d'y réaliser ma quatrième année de thèse. A Montréal, j'ai eu l'occasion de collaborer beaucoup plus avec Claudia L. Kleinman et Nicolas Rodrigue, ce qui nous a permis de publier un article sur une amélioration algorithmique du cadre probabiliste pour optimiser

les potentiels statistiques [Bonnard et al., 2009] (chapitre 5). S'appuyant sur cette amélioration, Claudia L. Kleinman s'est dirigée vers la création et l'optimisation de nouvelles formes de potentiels, tandis que de mon côté je me penchais sur le problème de l'inclusion de structures alternatives dans la procédure d'optimisation (chapitre 7). Le mémoire présenté ici représente les quatre années de travail de ma thèse, mais également une partie du travail que j'avais déjà eu l'occasion de faire à Montréal auparavant, et son résultat principal est le cadre statistique lui-même, qui est facilement adaptable à d'autres formes de potentiels, et qui peut être modifié afin d'inclure d'autres méthodes d'optimisation que l'on peut alors comparer aux méthodes déjà implémentées.. Cette thèse se trouvant à l'interface entre deux domaines de recherche très conséquents, j'ai essayé, dans le chapitre 1, de faire une introduction du modèle d'évolution de protéines soumis à des contraintes structurales. Bien que cette introduction ne soit pas exhaustive, et qu'il ne s'agisse pas du coeur de mes travaux, elle est nécessaire pour comprendre le cadre de travail et avoir une vision d'ensemble du projet. Dans un deuxième chapitre, j'ai fait une autre introduction, sur l'optimisation de potentiels statistiques, qui touche plus au coeur du sujet. Le troisième chapitre porte sur les méthodes numériques et statistiques que j'ai utilisé au cours de cette thèse. Le travail effectué pendant la thèse est décrit dans les quatre chapitres suivants. Le quatrième chapitre décrit le cadre probabiliste que nous avons développé, et le cinquième chapitre une amélioration algorithmique importante de la méthode. Nous nous sommes aperçus que la formulation présentée au quatrième chapitre était partiellement insatisfaisante, et le sixième chapitre permet la reformulation de la méthode dans un cadre plus consistant avec le modèle d'évolution sous-jacent. Le septième chapitre présente une méthode d'optimisation modifiée, en incluant des structures alternatives dans la procédure d'optimisation. Enfin, le dernier chapitre présente les améliorations futures et une conclusion sur mes travaux de thèse.

Il existe de nombreuses formes de potentiels statistiques, et de représentations structurales possibles. Cependant, les potentiels statistiques avec lesquels j'ai choisi de travailler ne sont constitués que de deux termes : un terme de contact (ou de distance discrétisée) entre deux types acides aminés et un terme d'accessibilité au solvant, également discrétisé. Afin de les obtenir, j'ai utilisé un programme de pré-traitement, implémenté par Claudia L. Kleinman, qui permet, à partir d'une structure de la Protein Data Bank, d'obtenir une matrice de contact entre deux sites, et une matrice d'accessibilité au solvant pour chaque site. Ensuite, pour chaque couple d'acide aminé (a, b), on peut définir un potentiel de contact ε_{ab}, et à chaque acide aminé a est associé un vecteur de valeurs α_a^d. Ces valeurs correspondent à l'énergie d'accessibilité au solvant pour chaque classe d pour cet acide aminé a.

Ces deux termes peuvent paraître très simples pour représenter des potentiels statistiques, mais ils sont très facilement généralisables à d'autres termes (comme par exemple un terme dépendant de la structure secondaire de la protéine). A l'aide de ces deux termes, je me suis donc consacrée à l'étude de différentes méthodes d'optimisation, tout en gardant à l'esprit qu'il faudrait par la suite intégrer des termes supplémentaires (comme ceux présentés dans [Kleinman et al., Submitted]).

Dans le cadre du modèle d'évolution, la structure est considérée comme constante le long de l'arbre pour toutes les structures protéiques. Une nouvelle fois, les données structurales sont constituées de la matrice de contact et de la matrice d'accessibilité au solvant. Même si les structures ne sont pas réellement constantes au cours du temps et le long des séquences, on peut supposer que les matrices utilisées comme définition structurale sont suffisamment grossières pour correspondre à la fois aux séquences actuelles et aux séquences ancestrales.

A l'aide de cette représentation simplifiée de la structure, et des potentiels statistiques associés, il est possible de comparer les résultats de différents modèles d'évolution dans un même cadre [Rodrigue et al., 2009]. Bien que les modèles SC implémentés dans ce cadre ne donnent pas encore de meilleurs résultats que certains autres modèles d'évolution, ils permettent d'obtenir un éclairage intéressant sur l'évolution moléculaire, et peuvent être utilisés dans certains domaines, comme par exemple dans la reconstruction des séquences ancestrales.

Introduction

> Scientific studies of evolution really started with Charles Darwin. He published his book *The Origin of Species* when he was fifty years old (Darwin, 1859), half a century after Lamarck's *Philosophie Zoologique*. With his masterly writing and wide ranging examples, Darwin not only persuaded the world that evolution has actually occurred, but also he showed through his theory of natural selection why adaptive evolution is an inevitable process. *The Origin of Species* has had an immeasurable influence not only on biology but also on human thought in general. We cherish Darwin for we owe to him our enlightened view of the nature of living things, including ourselves ; our civilization would be pityfully immature without the intellectual revolution led by Darwin, even if we are equally well off economically without it. H.J. Muller (1960), in celebrating the hundredth anniversary of the publication of *The Origin of Species* remarked that it can justly be considered as the greatest book ever written by one person.
>
> *M. Kimura* (1983) The neutral theory of molecular evolution

En cette année célébrant le 200e anniversaire de la naissance de Charles Darwin (1809-1882), et les 150 ans de l'Origine des espèces, Darwin et l'évolution font l'objet de nombreuses manifestations, à la fois pour les spécialistes du domaine, mais aussi pour les néophytes. Bien que faisant la part belle au naturaliste dont le livre a bouleversé notre vision du monde, le grand public oublie souvent les autres pères de ces théories qui ont donné corps au domaine de la phylogénie. Jean-Baptiste Lamark (1744-1829), notamment, qui publia son livre sur la théorie du transformisme (l'utilisation répétée de certaines caractéristiques conduit à une adaptation qui est transmise d'une génération à l'autre) en... 1809. Sa théorie fut réfutée par la suite, notamment par August Weismann (1833-1914), par l'observation que les modifications observées sur les cellules somatiques ne sont pas transmises à la génération suivante : ce rôle est dévolu aux cellules germinales. Cependant, les travaux de Lamark constituent une étape importante vers la théorie de l'évolution moderne.

Parallèlement à, et indépendamment de Darwin, Alfred Wallace (1823-1913), parvint à la même conclusion : l'évolution est le résultat de la *sélection naturelle*. Ainsi, lors de l'évolution, les caractères favorables sont conservés et les caractères défavorables sont éliminés : si un organisme est mieux adapté à son environnement qu'un autre, il tendra à plus se se reproduire que ce dernier.

Il manquait malheureusement aux travaux de Darwin une donnée essentielle : le support de l'information génétique, et donc le mécanisme par lequel les organismes évoluent. Darwin n'avait pas fait le lien avec les lois de Mendel (1822-1884) qui définissent les lois de la génétique, et comment se transmettent les caractères de génération en génération. Ce n'est que vers les années 1920-1930 que les conséquences de ces lois sont étudiées à l'échelle de la population, par Ronald Fisher (1890-1962), Sewall Wright (1889-1988) et John Haldane (1892-1964), à l'aide de modèles mathématiques. Liant les lois de Mendel aux travaux de Darwin, ils donnèrent naissance à la *génétique des populations*. L'idée fondamentale est que l'on peut ignorer les complexités liées aux variations de taille et de structure de population, ainsi qu'aux aléas environnementaux. En se focalisant alors simplement sur les taux de mutation, d'une part, et, d'autre part, sur la viabilité et la fertilité moyenne des individus ayant un allèle particulier en un locus d'intérêt, on peut prédire si cet allèle va disparaître, se fixer dans la population, ou se maintenir à une fréquence stable. Les conceptions initiales de la génétique des populations proposées, en particulier, par Fisher, étaient surtout déterministes : considérant que les populations naturelles sont très grandes, elles présupposaient qu'un mutant ne pouvait être fixé que s'il offrait un avantage sélectif.

Au cours des années 1960, cependant, il est apparu que les variations observées au niveau moléculaire étaient trop importantes pour pouvoir être expliquées uniquement par de la sélection positive. C'est dans ce contexte que Kimura proposa la théorie neutraliste. L'idée fondamentale est que, pour les populations de taille finie, des mutants neutres ou légèrement délétères peuvent être fixées, et que l'essentiel de la variation observée pourrait bien être de cette nature.

De manière un peu paradoxale, en apparence, la théorie neutraliste met en avant le principe de la sélection négative : les mutations qui ne sont pas fixées sont des mutations délétères, et parmi celles qui sont fixées, seule une infime fraction provient de la sélection positive. Dans la théorie de Kimura, on peut réécrire que le taux de substitution r de mutants dépend du taux de mutation, μ, et d'un facteur représentant la fraction des mutations neutres, f_0 [Kimura, 1983] :

$$r = f_0 \mu. \tag{1}$$

Une sélection entièrement neutre, tout en gardant un aspect de sélection positive, est cependant encore un peu trop restrictive pour expliquer certaines données biologiques. Aussi, en 1973, Ohta propose ce qu'on appellera par la suite le modèle quasi-neutre [Kimura and Ohta, 1974] : une mutation peut être acceptée si elle n'induit qu'un léger désavantage pour la protéine. Le problème du modèle proposé initialement par Kimura et Ohta est qu'il considère que seules les mutations neutres, ou légèrement délétères, peuvent être acceptées. Cette idée est problématique, car des substitutions exclusivement délétères résulteraient en un processus d'évolution vers des protéines de plus en plus délétères. En fait, on peut imaginer plus raisonnablement un modèle de sélection stabilisante : à l'équilibre, bien que la plupart des mutations proposées soient délétères, l'essentiel est éliminé par la sélection. Quand à celles qui parviennent à la fixation, elles sont soit légèrement délétères, soit légèrement avantageuses, de sorte que le bilan est équilibré.

Quantitativement, un tel modèle peut être formalisé [1], en s'appuyant sur une équation très simple, décrivant les relations entre substitution, mutation, taille de la population (N) et fixation :

$$R^{sub} = Q^{mut} \cdot 2N \cdot p^{fix}. \tag{2}$$

A noter que, dans le cas du modèle neutraliste strict, $p^{fix} = 1/2N$ et dès lors :

$$R^{sub} = \begin{cases} Q^{mut} \cdot 2N \cdot \frac{1}{2N} & \text{si la mutation est neutre (fraction } f_0 \text{ des mutations proposées)} \\ Q^{mut} \cdot 2N \cdot 0 & \text{si la mutation est délétère (fraction } 1 - f_0 \text{ des mutations)} \end{cases} \tag{3}$$

et donc, au bout du compte,

$$R^{sub} = f_0 \cdot Q^{mut}, \tag{4}$$

ce qui avait déjà été exprimé à l'équation (1).

Malgré toutes ses intuitions, Darwin ne connaissait pas le support (*génotype*) des caractères transmis aux générations suivantes (*phénotype*). A son époque, les biologistes ne pouvaient se baser que sur les caractères morphologiques pour retracer les relations entre les espèces (phylogénie). Le support de l'information génétique ne fut découvert que plus tard, d'abord par Thomas Morgan (1866-1945), puis par Oswald Avery (1877-1955). Le premier montra que les chromosomes sont un support de l'information génétique, et le second qu'il s'agissait plus précisément de l'ADN (acide désoxyribonucléique), composant ces chromosomes.

Cependant, il faudra attendre l'avènement de la biologie moléculaire, puis, en 1986, l'apparition de la PCR (*Polymerase Chain Reaction*) par Kary Mullis (1944-) pour que la production de données moléculaires prenne pleinement son essor. Cette méthode, qui

1. et le sera dans le chapitre 1.

permet d'amplifier un brin d'acide nucléique (ADN ou ARN [2]), et les progrès observés en biologie moléculaire [3] conduisent aujourd'hui à l'explosion des données moléculaires.

A partir de telles données moléculaires, il est possible de reconstruire l'histoire de la divergence des séquences génétiques entre les espèces. Les méthodes plus classiques comme la méthode du maximum de parcimonie ont petit à petit été remplacées par des méthodes probabilistes. Ces dernières sont particulièrement attrayantes, car elles permettent de formaliser les allers-retours entre les hypothèses du modèle et les estimations empiriques, mais également de spécifier des méthodes de comparaison de modèles. Le premier algorithme probabiliste efficace, introduit par Felsenstein [Felsenstein, 1981], dans le contexte du maximum de vraisemblance (ML), contribua particulièrement à la transition vers les méthodes probabilistes. Dans les méthodes ML, il s'agit de reconstruire la phylogénie la plus vraisemblable, au sens probabiliste du terme, c'est à dire que l'on cherche à maximiser la probabilité d'observer des données par rapport à la topologie de l'arbre. La méthode conduit au "meilleur" arbre possible, et, si l'on souhaite connaître le support statistique de cette phylogénie, il est possible de recourir à une méthode de *bootstrap*, qui consiste à générer un grand nombre d'arbres à partir de rééchantillonnage des données. On oppose souvent aux méthodes ML les méthodes Bayésiennes, introduites dans les années 1996-1998 par Yang et Rannala [Yang and Rannala, 1997], et Larget et Simon [Larget and Simon, 1999]. Les méthodes Bayésiennes consistent non pas à rechercher l'arbre optimal, en un quelconque sens, mais à échantillonner des arbres phylogénétiques de la distribution a posteriori, puis à calculer des moyennes sur ces échantillons.

A l'aide de telles méthodes de reconstructions, on peut distinguer deux orientations pour l'étude de l'évolution. La première, la phylogénie, consiste à reconstruire des arbres de parentés les plus fidèles possibles. Historiquement, il s'agit de la première approche de l'évolution (bien qu'au début il n'était possible de se baser que sur des caractères morphologiques), et bon nombre de méthodes ont été développées dans cette direction, que l'on peut séparer en trois grands groupes : les méthodes de distances, les méthodes de parcimonie et les méthodes probabilistes. Ces dernières ont été intensivement développées ces dernières années, et donnent des résultats de plus en plus précis, afin d'essayer de reconstruire des arbres fidèles à la fois dans les représentations des relations entre espèces (la *topologie* de l'arbre) mais aussi dans les distances évolutives (*longueurs de branches*) séparant ces espèces. Les méthodes de reconstruction font appel à des jeux de données de plus en plus conséquents, à la fois en terme de taille de séquences et de nombre d'espèces

2. Acide ribonucléique.

3. Une nouvelle méthode semble pouvoir permettre le séquençage d'un génome humain en quatre semaines [Pushkarev et al., 2009].

considérées, créant de nouveaux défis pour les différentes méthodes de reconstruction. Les arbres reconstruits sont considérés comme de plus en plus probants, mais il reste encore des écueils au sein de la reconstruction phylogénétique. Par exemple, les phylogénies peuvent être faussées par l'apparition de séquences analogues (semblables) mais non homologues (héritées d'un ancêtre commun) ou par le phénomène de l'attraction des longues branches (deux séquences évoluant rapidement sont considérées comme parentes alors qu'il n'en est rien).

La seconde orientation, dite de l'évolution moléculaire, s'oppose en partie à la phylogénie. Il ne s'agit pas ici de reconstruire l'arbre le plus fidèle possible, mais de comprendre le mécanisme de l'évolution. Les modèles d'évolution moléculaire visent à reproduire le phénomène évolutif, d'une manière mécanistique, en terme de mutation, de sélection, et de dérive génétique [4]. Là où un modèle éprouvé de reconstruction phylogénétique utiliserait une loi gamma pour représenter la variation des vitesses d'évolution entre les sites, un modèle d'évolution moléculaire définirait de manière explicite les relations entre les différents acteurs supposés causant cette variation de vitesse. Ainsi, si un modèle d'évolution moléculaire peut être moins efficace qu'un modèle phylogénétique, il permet néanmoins de tester les raisons supposées de la variation des vitesses d'évolution entre les sites.

Une distinction intéressante entre les différents modèles probabilistes peut être faite : les modèles empiriques (que l'on peut aussi appeler phénoménologiques [Rodrigue, 2007]), et les modèles mécanistiques. Les premiers cherchent à reconstruire des modèles qui correspondent le mieux aux données, sans chercher vraiment à décrire le processus sous-jacent, alors que les deuxièmes cherchent à développer des modèles expliquant le processus d'évolution (ici le processus de substitution) à l'aide de principes plus fondamentaux. Bien sûr, il existe tout une pléthore de modèles intermédiaires entre un modèle purement phénoménologique et un modèle entièrement mécanistique, mais d'une manière générale, les modèles utilisés en phylogénie sont des modèles plutôt empiriques, et les modèles d'évolution moléculaire s'orientent, par définition (et autant que faire se peut), vers les modèles mécanistiques.

C'est dans ce cadre que se place la thèse qui est présentée ici. Il s'agit à ce niveau, non pas de reconstruire des phylogénies précises, mais de tester différentes hypothèses mécanistiques afin de comprendre quelle part de l'évolution est expliquée par les différentes parties du modèle d'évolution proposé. Le modèle utilisé ici a été développé par Nicolas Rodrigue [Rodrigue et al., 2005, Rodrigue et al., 2006, Rodrigue et al., 2009, Rodrigue,

4. A cause des phénomènes aléatoires régissant les transmissions des allèles d'une génération à l'autre (par exemple par les rencontres entre individus lors d'une reproduction sexuée), les fréquences alléliques sont soumises à de grandes variations dans les petites populations.

2007], et s'articule autour de l'hypothèse que les protéines évoluent sous la contrainte de leur structure tridimensionnelle, selon la forme décrite à l'équation (2). La probabilité de fixation dépend notamment d'une fonction de score entre la séquence (qui subit la mutation) et la structure tridimensionnelle (qui représente ici une contrainte pour la fixation de la mutation). Ce modèle d'évolution, soumis à des contraintes structurales (SC) est un modèle hybride, à la fois mécanistique et phénoménologique : le phénomène de substitution est décrit d'une manière entièrement mécanistique, alors que la fonction de score structure/séquence est une fonction empirique (un potentiel statistique). Ce modèle a été implémenté de manière particulièrement raffinée [Rodrigue, 2007] dans un cadre bayésien, à l'aide de chaînes de Markov Monte Carlo.

Les séquences des protéines se replient dans l'espace, afin de former une structure stable thermodynamiquement. Les structures [Chothia and Lesk, 1986] évoluent lentement par rapport aux séquences. Ceci suggère que la structure incarne une contrainte (dont la forme exacte change très lentement au cours du temps) à laquelle la séquence doit à tout moment d'adapter. Vue sous cet angle, l'évolution peut être vue comme incarnant un principe d'optimisation analogue à un problème de *protein design*. De nombreuses études ont été faites pour explorer les relations entre la structure et la séquence, que ce soit par la création de champs de force semi-empiriques ou des potentiels statistiques, mais à ce jour, l'intégration explicite de la relation entre la structure et la séquence dans un modèle d'évolution est plutôt rare.

La reconstruction explicite du cheminement mutationnel entre deux séquences[5], supposé par le modèle d'évolution, permet notamment d'essayer de reconstituer les séquences ancestrales. Dans cette optique, les modèles SC sont particulièrement intéressants car la structure exerce une contrainte réelle, bien que incomplètement quantifiée, sur la séquence de la protéine, lors de l'évolution. En 2006, Williams et al comparèrent différentes méthodes de reconstructions, afin de déterminer quelle était la méthode la plus adaptée pour la reconstruction de séquences ancestrales [Williams et al., 2006]. A partir de séquences actuelles, ils faisaient évoluer les séquences en éliminant au fur et à mesure les séquences inadaptées (de part un repliement incomplet ou inexact). Ensuite, à partir des séquences générées, ils utilisaient trois méthodes de reconstruction basées sur les méthodes du maximum de parcimonie, du maximum de vraisemblance et une méthode Bayésienne. Connaissant le cheminement exact du modèle, il était alors facile de comparer les réelles séquences ancestrales avec les séquences ancestrales inférées par les différentes méthodes. La méthode du maximum de vraisemblance était celle qui fournissait de meilleures séquences ancestrales (suivit de peu par la méthode Bayésienne), mais il apparut également

5. séquences réelles (actuelles) ou inférées (aux noeuds internes de l'arbre).

que les séquences inférées par la méthode du maximum de vraisemblance étaient trop stables thermodynamiquement par rapport aux séquences réelles, alors que la méthode Bayésienne semblait, dans ce contexte bien précis, être plus adaptée que les deux autres méthodes, en ce qu'elle résultait en des séquences marginalement stables, à l'instar des protéines naturelles.

Des modèles SC peuvent également être utilisés pour la prédiction de mutants délétères, par exemple. En 2005, Stone et Sidow, montraient que les propriétés physico-chimiques des acides aminés pouvaient être utilisées pour prédire si une mutation était délétère [Stone and Sidow, 2005]. A partir d'alignement de séquences et d'une phylogénie, ils déterminaient les profils physico-chimiques de chaque site de la protéine. Ils purent alors montrer que les différents degrés de maladie corrélaient avec la violation de ce profil physico-chimique. Il serait intéressant de voir, en utilisant un modèle SC en lieu et place du modèle phylogénétique utilisé, si la prédiction serait meilleure (et si oui, de combien serait l'amélioration).

Cependant, le modèle d'évolution soumis à des contraintes structurales impose que le lien entre la séquence et la structure soit clairement déterminé : tout le modèle repose sur ce lien. Comme on vient de l'évoquer, il existe déjà plusieurs méthodes permettant de caractériser, de manière plus ou moins précise, les relations entre une séquence protéique et son repliement tridimensionnel. Cependant, le cadre du modèle d'évolution est déjà très demandeur en ressources statistique (modèle probabiliste lourd), biologique (taille des jeux de données) et informatique (temps de calcul) ; imposant que la relation structure/séquence puisse être définie de manière simple. Nous avons donc choisi de représenter cette relation structure/séquence à l'aide de potentiels statistiques. Une question s'est alors posée : quel potentiel statistique utiliser ? On peut par exemple se demander si un potentiel optimisé dans un contexte de *protein folding* serait adapté à une utilisation dans un modèle d'évolution dont les termes sont plus proches d'un problème de *protein design*[6]. Au delà, l'idée est aussi de construire un cadre probabiliste permettant de comparer différentes méthodes de construction de potentiels, afin de tester leur efficacité dans le contexte du modèle d'évolution SC.

Cette thèse a pour objet de répondre au besoin, exprimé par le modèle phylogénétique, d'une fonction simple prenant la forme d'un potentiel statistique, reliant la structure et la séquence d'une protéine, définie dans un cadre entièrement probabiliste[7] et adapté à la perspective évolutive. Il est important de garder à l'esprit que les paramètres de cette

6. Nous en parlerons plus en détail dans le chapitre 2.

7. Il s'agit ici de garder une cohérence entre le modèle d'évolution, probabiliste, et les relations structure/séquence.

fonction seront toujours optimisés dans l'optique d'être intégrés dans ce modèle SC et que s'ils peuvent trouver des applications dans d'autres domaines, cela n'est pas leur but premier.

Le premier chapitre de ce mémoire est consacré à une description non exhaustive du cheminement intellectuel menant au modèle d'évolution SC. En effet, si ce modèle ne fait pas vraiment partie des méthodes développées ici, il représente à la fois la raison d'être et le cadre dans lequel seront testés les potentiels statistiques optimisés et il est donc important d'appréhender également ce domaine. Le deuxième chapitre est dédié à la description de la fonction représentant les liens structure/séquence et aux méthodes d'optimisations utilisées pour la construction des potentiels statistiques. Le troisième chapitre correspond à une description des méthodes statistiques et numériques utilisées dans cette thèse. Les quatre chapitres suivants sont principalement constitués par les articles publiés ou en cours de soumission. Le quatrième chapitre correspond à une première définition du cadre probabiliste de la méthode d'optimisation. Le cinquième chapitre représente une amélioration algorithmique non négligeable de la méthode. Cependant, la définition proposée au chapitre 4 est partiellement insatisfaisante, et le sixième chapitre se consacre donc à la reformulation de la méthode de manière plus consistante avec le cadre phylogénétique. Le septième chapitre expose une nouvelle amélioration de la méthode, par l'intégration de structures leurres dans la procédure d'optimisation et dans le modèle SC. Ces trois chapitres représentent un travail exclusivement effectué pendant la thèse. Le dernier chapitre relate les améliorations qui peuvent être proposées et rapidement mises en œuvre, ainsi que les perspectives de cette thèse.

Première partie

Etat de l'art

Chapitre 1

Modèles d'évolution probabilistes

1.1 Introduction

La phylogénie et l'évolution moléculaire se consacrent à l'étude des relations inter-espèces (ou inter-gènes) pour reconstruire l'histoire de la divergence entre les espèces (ou entre les gènes). Cependant, si en phylogénie l'on s'attache à reconstruire des arbres de parentés les plus fidèles possibles, l'on préfèrera en évolution moléculaire décrire le processus d'évolution de manière mécanistique. De ce fait, il s'agit plus de comprendre le processus évolutif (et de tester différentes hypothèses évolutives), que de retrouver une phylogénie exacte. Cela est réalisé en modélisant le processus de variation génétique sur la longue durée. Ces variations sont le résultat de processus complexes à petite échelle évolutive et qui sont l'objet d'étude de la génétique des populations. Elles sont essentiellement causées par des mutations ponctuelles, même si l'on observe aussi des phénomènes à grande échelle, comme des duplications entières de gènes. En ne considérant que les phénomènes à petite échelle, trois types d'évènements peuvent causer les variations observées dans les séquences codantes : l'insertion, la délétion et le remplacement d'un nucléotide par un autre, mais nous ne considèrerons ici que ce dernier cas de figure.

Les modèles d'évolution phénoménologiques sont construits de manière à modéliser directement le processus de substitution, en considérant l'ensemble du processus, qui va de la mutation à la fixation du mutant dans la population, comme un évènement ponctuel dans le temps. Par la suite, on fera bien la distinction entre *mutations* (au niveau des individus) et *substitutions* (au niveau de la population). On peut modéliser le processus de substitution selon un modèle Markovien, caractérisé par une matrice. Cette matrice (que l'on notera Q) contient toutes les informations qui permettent de calculer les probabilités de substitution d'un nucléotide/acide aminé/codon en un autre. La matrice de Dayhoff

Nom	Code à trois lettres	Code à une lettre
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cystéine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Méthionine	Met	M
Phénylalanine	Phe	F
Proline	Pro	P
Sérine	Ser	S
Thréonine	Thr	T
Tryptophane	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

TABLE 1.1 – Alphabet standard des acides aminés.

sur les substitutions dans une séquence protéique [Dayhoff et al., 1972] est un excellent exemple d'un modèle phénoménologique : on observe qu'un acide aminé se substitue plus souvent en un autre, mais on ne cherche pas à séparer les contributions respectives de la mutation et de la sélection.

Dans ce mémoire, je ne parlerai que des méthodes probabilistes, c'est à dire des méthodes du maximum de vraisemblance et des méthodes Bayésiennes [8]. Je commencerai par décrire les principes généraux des modèles Markovien de substitution, avant de revenir un peu plus en détail sur les modèles d'évolution nucléotidiques, à acides aminés et à codons, avant de parler plus spécifiquement des modèles d'évolution mutation/sélection.

Deux types de séquences seront considérés dans ce chapitre. Les premières sont les séquences nucléotidiques, d'ADN ou d'ARN. Ces séquences, que nous noterons σ, sont formées à l'aide de quatre nucléotides : A (adénine), C (cytosine), G (guanine), T (thymine) pour l'ADN et A, C, G, U (uracile) pour l'ARN [9]. Les séquences protéiques, notées s, sont composées à partir de l'alphabet des vingt acides aminés standards, présentés dans la table 1.1 [10].

1.2 Modèles de substitutions

Dans cette section, nous noterons σ la séquence telle que $\sigma_i = a$, et σ' la séquence telle que $\sigma'_i = b$. Ces deux séquences sont plus proches voisines (c'est à dire qu'elles ne diffèrent qu'en la position i) :

$$\forall j \neq i \quad \sigma_j = \sigma'_j. \tag{1.1}$$

1.2.1 Modèle de substitution markovien

Dans leur grande majorité, les modèles d'évolution probabilistes supposent que l'évolution d'une séquence en une autre obéit à un *processus markovien*, "sans mémoire". Ainsi, à un moment donné, l'évolution de la séquence σ vers la séquence σ' ne dépend que de σ et aucunement des séquences qui la précèdent. Les modèles d'évolution les plus simples admettent généralement quatre hypothèses supplémentaires au modèle de substitution

8. Même s'il existe d'autres méthodes (maximum de parcimonie et méthodes de distances) nous ne les évoquerons pas.

9. Lors de transcription de l'ADN en ARN, la thymine présente dans l'ADN est remplacée par l'uracile dans l'ARN.

10. Bien qu'il existe deux autres acides aminés qui ne font pas partie de cet alphabet standard, ils ne seront pas considérés ici.

markovien : la stationnarité, la réversibilité, l'indépendance entre positions et l'homogénéité au cours du temps. Bien qu'il existe des modèles qui permettent de relaxer une ou plusieurs de ces hypothèses (car elles ne sont pas forcément compatibles avec des propriétés biologiques observées), elles sont souvent préférées à des hypothèses plus réalistes afin de limiter la complexité du modèle.

Tout d'abord, le processus de substitution est supposé *stationnaire*, à l'équilibre, et donc que la probabilité d'observer b à une position donnée est toujours égale à sa probabilité stationnaire, π_b. Soit ρ_{ab} le taux d'échange de a vers b dans la chaîne de Markov. En imposant :

$$\rho_{ab} = \rho_{ba}, \tag{1.2}$$

le processus markovien devient *réversible*, c'est à dire que la probabilité d'observer un échange de a vers b est la même que celle d'observer un échange de b vers a :

$$\pi_a P_{ab}(t|\mathbf{Q}) = \pi_b P_{ba}(t|\mathbf{Q}), \tag{1.3}$$

où $P_{ab}(t|\mathbf{Q})$ représente la probabilité de substituer b à a en un temps t, sachant la matrice de substitution $\mathbf{Q} = [Q_{ab}]$. Si l'on pose que le processus de Markov est réversible grâce à l'équation (1.2), cette matrice de substitution peut être définie de la manière suivante :

$$Q_{ab} = \begin{cases} \pi_b \rho_{ab} & \text{si } a \neq b \\ -\sum_{c \neq a} Q_{ac} & \text{si } a = b \ , \end{cases} \tag{1.4}$$

Ce modèle est appelé modèle GTR (*General Time reversible*), et correspond au modèle de substitution le plus général. La paramétrisation décrite dans l'équation (1.4) permet d'introduire directement les probabilités stationnaires dans l'expression de $\mathbf{Q}$, ce qui est notamment utile pour le calcul de la vraisemblance, à la base des modèles probabilistes.

Le mécanisme de substitution en chaque site est considéré comme *indépendant* des autres sites. Si cette hypothèse n'est pas le reflet d'observations biologiques (comme nous en parlerons dans la section 1.3.2), elle est extrêmement pratique car elle permet principalement de factoriser les calculs.

On suppose généralement que le processus est *homogène* (le modèle de substitution est le même le long de la séquence et le long de l'arbre) et que donc la probabilité de substitution d'un nucléotide en un autre est la même quel que soit le site considéré. S. Blanquart propose une définition intéressante de cette hypothèse d'homogénéité en séparant *l'homogénéité qualitative* (le processus est le même) de *l'homogénéité quantitative* (la vitesse de substitution est la même entre les sites et au cours du temps) [Blanquart, 2007]. Très tôt, la relaxation de l'hypothèse d'homogénéité quantitative le long de la

séquence, afin d'autoriser des vitesses de substitution différentes entre les sites, a été rendue possible par l'introduction de la loi gamma [Yang, 1993], qui définit la distribution des vitesses de substitution entre les sites [11]. L'hypothèse de l'homogénéité quantitative des sites au cours du temps (ou *horloge moléculaire*) a été naturellement relaxée par la définition de longueurs de branches variables, arbitraires, sans aucune contrainte visant à rendre l'arbre ultramétrique.

L'homogénéité qualitative suppose que les taux d'échanges et les fréquences stationnaires (donc les préférences pour tel nucléotide ou tel acide aminé) sont les mêmes au cours du temps et entre les sites [12]. Cette hypothèse n'a été relaxée que récemment par l'utilisation de modèles de mélange [Thorne et al., 1996, Koshi and Goldstein, 1997, Goldman and Whelan, 2002, Lartillot and Philippe, 2004, Le et al., 2008a], qui définissent un modèle de substitution général, avec des matrices de substitution différentes pour les différentes catégories de sites.

1.2.2 Vraisemblance

Soit $\mathbf{P}(t|\mathbf{Q}) = [P_{ab}(t|\mathbf{Q})]$ la matrice représentant les probabilités de substitution. Étant donné une matrice de substitution $\mathbf{Q}$, on peut définir la matrice $\mathbf{P}(t|\mathbf{Q})$ telle que :

$$\mathbf{P}(t|\mathbf{Q}) = e^{\mathbf{Q}t}, \tag{1.5}$$

$e^{\mathbf{Q}t}$ est une matrice, appelée exponentielle de la matrice $\mathbf{Q}t$, c'est à dire que :

$$e^{\mathbf{Q}t} = \sum_{n \geq 0} \frac{(\mathbf{Q}t)^n}{n!}. \tag{1.6}$$

Si σ et σ' sont des séquences à deux nœuds successifs le long d'un arbre, alors t correspond à une longueur de branche λ. On caractérisera une phylogénie par sa topologie $\mathcal{T}$ et ses longueurs de branche Λ, et on supposera que σ et σ' sont situées à deux nœuds consécutifs d'un arbre, et donc séparés par la *distance évolutive* λ.

En calculant les probabilités en temps fini (comme donné par l'équation (1.5)), le long d'une phylogénie, et en sommant par programmation dynamique (algorithme de pruning ou de peeling [Felsenstein, 1981]), il est alors possible de calculer $p(D_i|\mathcal{T}, \Lambda)$, où D_i représente les données au site i. Cette probabilité est une somme sur toutes les reconstructions ancestrales possibles. En faisant le produit sur tous les sites, on obtient :

$$p(D|\mathcal{T}, \Lambda) = \prod_{1 \leq i \leq n} p(D_i|\mathcal{T}, \Lambda). \tag{1.7}$$

11. On préfèrera en général utiliser une loi gamma discrétisée.

12. La relaxation des hypothèses d'homogénéité ne se posera pas dans le cas des modèles SC, puisqu'elle sera prise en compte intrinsèquement par le terme lié à la sélection (cf. 1.3.1).

Ce terme est appelé *vraisemblance* et est à la base de la méthode du maximum de vraisemblance et de la méthode Bayésienne.

La méthode dite "du maximum de vraisemblance" (ML) vise alors à maximiser cette vraisemblance [Felsenstein, 1981], c'est à dire que l'on va explorer l'espace des paramètres afin de trouver les valeurs qui rendent les données les plus probables possibles.

La méthode Bayésienne, quant à elle, cherche à estimer les paramètres du modèle par le calcul de leur probabilité *a posteriori*. Celle-ci est définie à partir de la probabilité *a priori* des paramètres (définie avant d'avoir vu les données), et de la vraisemblance.

Soit θ un jeu de paramètres tels que $\theta \in \Theta$. La probabilité a posteriori de ces paramètres est obtenue à l'aide du théorème de Bayes :

$$p(\theta|D, \mathcal{T}, \Lambda) = \frac{p(D|\theta, \mathcal{T}, \Lambda)p(\theta|\mathcal{T}, \Lambda)}{p(D|\mathcal{T}, \Lambda)}. \tag{1.8}$$

$p(\theta|\mathcal{T}, \Lambda)$ est la distribution *a priori* des paramètres et $p(D|\mathcal{T}, \Lambda)$ est la *vraisemblance marginale* :

$$p(D|\mathcal{T}, \Lambda) = \int_{\Theta} p(D|\theta, \mathcal{T}, \Lambda)p(\theta|\mathcal{T}, \Lambda)d\theta. \tag{1.9}$$

Dans le cas de la reconstruction phylogénétique, l'on s'intéresse à des espérances a posteriori sur $p(\theta|D, \mathcal{T}, \Lambda)$. Même pour des modèles relativement simples, de telles espérances sont incalculables analytiquement, et il faut contourner ce problème par l'utilisation de Chaînes de Markov Monte Carlo (MCMC). Une des méthodes que l'on peut utiliser pour estimer ces espérances est décrite dans le chapitre 3, avec les autres méthodes numériques auxquelles cette thèse fait appel.

1.2.3 Modèles d'évolution nucléotidiques

Comme indiqué précédemment, un modèle d'évolution moléculaire repose sur l'expression de sa matrice de substitution. Cette matrice représente le mécanisme intrinsèque du modèle, définissant les paramètres qui permettent à une séquence de se substituer à une autre, selon un processus markovien (cf. 1.2.1). On peut séparer les modèles d'évolution en trois catégories, les modèles nucléotidiques, les modèles à acides aminés et les modèles à codons. Les deux premiers correspondent la plupart du temps à des modèles phénoménologiques, alors que le dernier est souvent utilisé comme base pour les modèles mécanistiques.

Les premiers modèles développés utilisaient des matrices de mutations entre nucléotides, dont la forme est un cas particulier de l'équation (1.4). Le modèle le plus simple a été défini par Jukes et Cantor :

$$Q = \begin{Bmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{Bmatrix} \tag{1.10}$$

où la probabilité de substitution d'un nucléotide en l'un des trois autres est identique quels que soient les nucléotides [Jukes and Cantor, 1969]. La matrice de mutation de Jukes et Cantor est ici présentée sous sa forme normalisée [Bryant et al., 2005]. Cette matrice de mutation normalisée permet d'avoir un taux général de mutation non redondante égal à 1, ce qui élimine donc les possibilités de mutation d'un acide aminé à lui-même. Elle est obtenue en divisant la matrice initiale de Jukes et Cantor par $\mu = 3/4$. D'une manière plus générale, il est possible d'obtenir une matrice de substitution normalisée en divisant la matrice initiale par $\mu = \sum_a \pi_a Q_{aa}$.

Le modèle de Jukes et Cantor s'est vite avéré trop simpliste, et des améliorations ont été apportées aux modèles nucléotidiques, comme par exemple l'inclusion de taux de transversion, représentant la substitution entre une purine (A,G) et une pyrimidine (T,C), et de transition différents. Des fréquences stationnaires distinctes pour les quatre nucléotides ont également été introduites, pour finalement mener au modèle le plus général, spécifié par l'équation (1.4), également appelé modèle GTR (*general time reversible*) [Lanave et al., 1984]. Cependant, aussi élaborés soient-ils, les modèles purement nucléotidiques n'arrivent pas à traiter correctement les dépendances entre les positions d'un même codon.

1.2.4 Modèles d'évolution à acides aminés

Dans le cas des séquences codantes, la sélection agit principalement non pas sur la séquence nucléotidique codante, mais sur la séquence protéique codée, ce qui induit des dépendances entre les positions d'un même codon. Le code génétique standard (table 1.2) décrit la correspondance d'un triplet de nucléotides (*codon*) avec un acide aminé ou avec un arrêt de la traduction (codons *stop*). Avec les 4 nucléotides, il existe donc 64 codons différents, dont 3 correspondent à des codons stops (TAG, TGA et TAA) dans le code génétique standard. Les 61 codons restants codent pour les 20 acides aminés. Le code est donc redondant et deux codons correspondant au même acide aminé sont appelés des *codons synonymes*.

Les modèles purement nucléotidiques ne prenant pas en compte la séquence protéique codée, des modèles dits "à acides aminés" ont alors fait leur apparition. Ils permettent de considérer le phénomène de substitution au niveau de la protéine codée, soumise aux

	T		C		A		G		
T	TTT	Phénylalanine	TTC	Sérine	TAT	Tyrosine	TGT	Cystéine	T
	TTC		TCC		TAC		TGC		C
	TTA	Leucine	TCA		TAA	Stop	TGA	Stop	A
	TTG		TCG		TAG		TGG		G
C	CTT		CCT	Proline	CAT	Histidine	CGT	Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Isoleucine	ACT	Thréonine	AAT	Asparagine	AGT	Sérine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lysine	AGA	Arginine	A
	ATG	Methionine	ACG		AAG		AGG		G
G	GTT	Valine	GCT	Alanine	GAT	Aspartate	GGT	Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glutamate	GGA		A
	GTG		GCG		GAG		GGG		G

TABLE 1.2 – Le code génétique standard. La première colonne indique le premier nucléotide du codon, la première ligne indique le deuxième nucléotide du codon, et la dernière colonne indique le troisième nucléotide du codon.

contraintes sélectives, avec une complexité raisonnable. La matrice de substitution est alors écrite en fonction des acides aminés et est donc de taille 20×20. Un des modèles les plus simples (modèles poissons) considère que les taux d'échanges sont les mêmes quels que soient les deux acides aminés et que donc $\rho_{ab} = 1$. Ainsi,

$$Q_{ab} = \{\pi_b\} \forall\ a, b \in \{1..20\}, \tag{1.11}$$

correspond à la matrice de substitution du modèle F81 [Felsenstein, 1981].

Cependant, il existe une sélection purificatrice extrêmement forte au niveau de la protéine, puisque celle-ci doit pouvoir conserver sa fonction. Si une telle contrainte est évidente, la façon de la modéliser l'est beaucoup moins : quelle formulation du modèle permettrait le mieux de capter les effets de cette sélection au niveau des protéines ? Et comment le décrire de manière à ce que le problème reste abordable ?

La première approche [Dayhoff et al., 1972], consiste à dire que les mutations conservant les propriétés biochimiques des acides aminés sont les moins délétères et donc les plus susceptibles d'être fixées. Par exemple, le remplacement d'une valine par une alanine serait moins problématique pour la protéine que le remplacement de cette même valine par une arginine. Il s'agit donc de construire une matrice de substitution 20×20 de type GTR telles que les substitutions conservatrices soient les plus fréquentes. En pratique, ces paramètres sont appris sur les données. La première matrice [Dayhoff et al., 1972] utilisait des alignements de séquences afin d'estimer les paramètres de la matrice 20×20, en comptant le nombre de fois où des substitutions avaient été inférées entre les paires de séquences ayant 85 % d'homologie [13]. Ce type de matrice fut longtemps utilisé (et l'est encore souvent) dans les modèles d'inférences, dans sa version originale ou modifiée [Jones et al., 1992b] [14]. D'autres méthodes furent utilisées afin d'estimer des paramètres d'échange semblables, comme par exemple la très connue matrice WAG [Whelan and Goldman, 2001] ou LG [Le and Gascuel, 2008] dont les paramètres sont optimisés à l'aide de la méthode de maximum de vraisemblance.

Toutefois, si de telles matrices permettent de prendre en compte une sélection au niveau de la protéine, elles présentent le désavantage de représenter de la même manière les substitutions d'un acide aminé à un autre, quel que soit le site considéré (hypothèse d'homogénéité qualitative, voir ci dessus). Or, les contraintes liées à la sélection ne sont pas les

13. Ils supposaient qu'avec 85 % d'homologie, les différences entre les deux séquences étaient dues à des substitutions simples et non à des substitutions multiples (substitution de a vers b via d'autres acides aminés) [Jones et al., 1992b].

14. La méthode de calcul des paramètres d'échange reste la même, mais le choix des séquences sur lesquelles est effectué le comptage varie.

mêmes sur tous les sites de la protéine, et elles sont au contraire fortement hétérogènes : un acide aminé appartenant au site fonctionnel de la protéine n'a pas les mêmes propensions substitutionnelles qu'un acide aminé dont le seul rôle serait d'être à la surface de la protéine. Cependant, si l'hypothèse d'homogénéité entre sites ne peut pas être conservée, l'hypothèse d'une homogénéité dans le temps (la vitesse d'évolution est la même le long de l'arbre, pour un site donné), semble a première vue plus raisonnable [15].

Afin de représenter la variabilité entre les sites, la première approche proposée fut de conserver la matrice de substitution entre les acides aminés mais de modéliser des vitesses d'évolution différentes entre les sites à l'aide de la loi gamma. Cette loi gamma permet de déterminer la distribution des vitesses d'évolution entre les sites, et est généralement utilisée dans une forme discrétisée à quatre catégories [Yang, 1993].

Plus complexes, les modèles de mélange [Koshi and Goldstein, 1997, Goldman and Whelan, 2002, Lartillot and Philippe, 2004, Le et al., 2008a, Le et al., 2008b] permettent de déterminer des matrices du substitution selon la catégorie du site considéré. Cependant, si les matrices de substitution diffèrent selon la catégorie du site, le modèle reste formellement identiquement distribué (dans la mesure où la vraisemblance en chaque site est une moyenne pondérée sur toutes les catégories disponibles). Pour la détermination des catégories des sites, on peut soit fixer la catégorie de chaque site soit optimiser cette détermination en fonction des données. Pour choisir comment fixer les classes, il faut tout d'abord définir le critère de classification. Par exemple, Thorne et al essayèrent de relier la structure secondaire et les substitutions des acides aminés, en déterminant, l'organisation de la structure secondaire le long de la séquence [16], puis en attribuant un processus d'évolution pour chaque catégorie de site [Thorne et al., 1996]. D'un autre côté, on peut laisser la détermination du nombre de classes et de l'appartenance de chaque site à une classe à la seule appréciation des données [Lartillot and Philippe, 2004, Wang et al., 2008].

Cependant, ce type d'approche serait plutôt de type phénoménologique, tendant à modéliser correctement les paramètres sans pour autant expliquer les phénomènes sous-jacents, et encore moins de quantifier leur impact réel sur le modèle d'évolution. De plus, les matrices de substitution définies au niveau des acides aminés perdent de l'information, à cause des mutations synonymes, qui apparaissent dans la séquence nucléotidique, mais qui sont invisibles dans la séquence protéique. En outre, le code génétique (table 1.2) a un impact réel sur le processus de remplacement des acides aminés : une mutation entre deux acides aminés est plus susceptible d'être proposée si les codons menant à ceux-ci ne

15. Cette hypothèse est cependant remise en cause par différents travaux [Foster, 2004, Blanquart, 2007].

16. Ils supposaient que la structure secondaire était la même pour chaque site aligné, et déterminaient la structure secondaire d'une des protéines alignées.

diffèrent que d'un nucléotide que s'ils diffèrent de deux ou trois nucléotides. Par exemple, il semble plus probable qu'une lysine (AAA ou AAG) mute en une asparagine (AAT, AAC) qu'en une cystéine (TGT, TGC). Pour décrire une telle dépendance au niveau du code génétique sans perdre le lien à la séquence protéique, l'alternative la plus convaincante est de faire appel à des modèles à codons.

1.2.5 Modèles d'évolution à codons

Le but premier du modèle à codons présenté par Muse and Gaut était de paramétrer explicitement la différence entre les codons synonymes et non synonymes [Muse and Gaut, 1994]. Même s'il faudrait théoriquement définir une matrice de substitution de taille 64×64, les substitutions entre les codons stop et les codons menant à des acides aminés sont problématiques. En effet, des codons stop introduisent des variations dans la longueur de la séquence codée, de telles mutations sont trop délétères pour être fixées et on peut donc ignorer les codons stops dans le processus de substitution. Déjà, dans le modèle de Muse et Gaut, le processus de substitution n'était décrit qu'entre les 61 codons codants, et nous utiliserons également cette convention.

Par la suite, nous emploierons les notations simplifiées suivantes : supposons deux codons, c et c', plus proches voisins, c'est à dire qu'ils ne diffèrent qu'en une seule position. Notons b et b' les nucléotides observés en cette position dans les codons c et c' et a (resp. a') l'acide aminé codé par le codon c (resp. c'). On considèrera également que, si c et c' ne sont pas plus proches voisins, alors $Q_{cc'} = 0$. Alors, le modèle de substitution peut être décrit par la matrice de substitution $Q_{cc'}$:

$$Q_{cc'} = \pi_{b'} \rho_{bb'}, \tag{1.12}$$

où $\pi_{b'}$ est donc la fréquence stationnaire du nucléotide b' et $\rho_{bb'}$ le taux d'échange de b vers b'. La matrice de substitution, $Q_{cc'}$ est une matrice 61×61, et il est intéressant de noter que dans le modèle décrit par l'équation (1.12), le taux de substitution instantané de dépend pas de la fréquence stationnaire du codon, mais de celle du nucléotide muté. Ce modèle est noté $MG - F1 \times 4$ car il est basé sur le modèle originel de Muse et Gaut, et que la matrice des fréquences stationnaires est une matrice de dimension 1×4 (chaque nucléotide a une fréquence qui lui est propre, mais qui est la même quelle que soit sa position dans le codon).

Le modèle ci-dessus (eq. (1.12)) est un modèle purement neutre, mais dans un sens extrême : aucun effet sélectif n'est modélisé, à part la sélection purificatrice contre les mutation non-sens (contre l'apparition des codons stops). Pour prendre en compte les

effets sélectifs sur la séquence protéique, le modèle peut être modifié de la manière suivante [Muse and Gaut, 1994] :

$$Q_{cc'} = \begin{cases} \pi_{b'}\rho_{bb'} & \text{si } a = a' \\ \omega\pi_{b'}\rho_{bb'} & \text{si } a \neq a' \\ 0 & \text{si } c \text{ et } c' \text{ ne sont pas plus proches voisins} \end{cases} . \tag{1.13}$$

Le paramètre ω représente le ratio de substitutions non synonymes relativement aux substitutions synonymes ($\omega = dN/dS$). Son interprétation est exprimée en terme de sélection : à supposer que les substitutions synonymes sont neutres, alors $\omega < 1$ (resp. $\omega > 1$) signifie que, en moyenne, les substitutions non synonymes sont sous une sélection négative (resp. positive) [Yang and Bielawski, 2000]. Évidemment, tous les sites ne subissent pas le même type de sélection et de plus, la pression de sélection varie au cours du temps [Ross and Rodrigo, 2002]. Ainsi, des modèles ont été construits afin de faire varier ω entre les différents sites [Nielsen and Yang, 1998], mais également entre les branches [Guindon et al., 2004].

Le modèle présenté à l'équation (1.13) peut être modifié afin de prendre en compte les différentes modalités d'évolution selon les positions au sein d'un codon, en définissant des probabilités stationnaires différentes selon la position du nucléotide cible dans le codon. Le modèle ainsi créé est référencé comme le modèle $MG - F3 \times 4$ (puisque la matrice des fréquences stationnaires est de taille 3×4). Pour aller plus loin, il est également possible d'introduire des fréquences stationnaires définies par codon, et ne dépendant donc plus uniquement du nucléotide "cible". Ces modèles à codons 61×61 (le plus connu étant le modèle $GY - F61$) permettent notamment d'introduire de manière déguisée des taux de substitution différents entre les trois positions, des biais d'usage des codons ou des effets sélectifs au niveau codant [Goldman and Yang, 1994].

Au delà de ces modèles MG et GY, il existe également des modèles à codons purement empirique [Kosiol et al., 2007]. Alors que les modèles classiques n'autorisent des mutations que pour un seul site nucléotidique à la fois, ce modèle (appelé ECM) autorise les substitutions multiples (deux et trois nucléotides) au sein d'un même codon. Bien que l'apparition de substitutions multiples soit sujet à débat [Whelan and Goldman, 2004], Kosiol et al montrèrent que leur modèle empirique était significativement plus performant que les autres modèles à codons et remettant donc en question les modèles de type MG et GY.

Cependant, si l'introduction de ce genre d'effets complexes permet d'améliorer les résultats du modèle, il est probable qu'au moins une partie de ce qui est capté par ces

nouveaux paramètres soit de nature sélective, ce qui rend plus problématique l'interprétation de ω (qui était censé justement capter les effets sélectifs au niveau des codons non synonymes).

Plus fondamentalement, un paramètre ω tel que défini dans l'équation (1.13), même variable entre sites ou branches, ne fait pas de différence entre les différents types de substitutions non synonymes. Et donc, ainsi que nous le verrons dans la section 1.3.1, nous essayerons d'exprimer les différences entre les substitutions non synonymes autrement que par la simple utilisation de la matrice nucléotidique.

1.3 Modèles mécanistiques

Le principe des modèles à codons offre des possibilités bien au delà des simples modèles à ω, et qui commencent depuis peu à être explorées. En particulier, les modèles à codons permettent la formulation explicite de modèles mécanistiques, formulés en fonction des trois forces fondamentales de mutation, de sélection et de dérive génétique. Dans de tels modèles, les phénomènes de mutation et de sélection sont appliqués séparément sur les séquences qui y sont soumises (mutation sur la séquence nucléotidique et sélection sur la séquence protéique), mais sont reliées à l'aide de la matrice de substitution, faisant intervenir la dérive génétique via une probabilité de fixation dépendante de la taille de la population. Dans la littérature, ces modèles sont plus simplement appelés mutation/sélection.

1.3.1 Modèle d'évolution mutation/sélection

Les modèles dit de *mutation/sélection* prennent explicitement en compte un facteur mutationnel sur les séquences nucléotidiques et un facteur sélectif sur les séquences protéiques codées [17], ce qui permet de séparer ces deux facteurs bien distincts de l'évolution. La matrice de substitution que nous avons vu dans les sections précédentes devient dans de tels modèles la matrice de mutation, permettant de déterminer les propositions de mutation au sein de la séquence. Afin de différencier les deux matrices, nous noterons $\mathbf{R}$ la matrice de substitution, qui prendra donc en compte le facteur sélectif, et $\mathbf{Q}^{mut}$ correspondra désormais à la matrice de mutation.

Le taux de substitution d'un codon c vers un codon c' est exprimé de la manière suivante :

17. Ce facteur sélectif peut s'appliquer à d'autres niveaux, mais nous ne parlerons ici que de la sélection sur les séquences protéiques codées.

$$R_{cc'} = \begin{cases} Q^{mut}_{bb'} \cdot 2N \cdot p^{fix}(aa') & \text{si la mutation n'est pas synonyme} \\ Q^{mut}_{bb'} & \text{si la mutation est synonyme} \\ 0 & \text{si } c \text{ et } c' \text{ ne sont pas plus proches voisins} \end{cases}, \quad (1.14)$$

où N représente la taille de la population. Le terme de mutation, $Q^{mut}_{bb'}$ s'applique uniquement sur les nucléotides du codon, alors que la sélection, représentée par la probabilité de fixation $p^{fix}(aa')$, ne dépend que des acides aminés codés. Les modèles testés peuvent être d'une grande variété, et l'on peut ainsi essayer de déterminer, avec un même modèle de mutation, différents types de sélection.

On peut particulièrement utiliser les développements issus de l'application de la théorie de la diffusion à la génétique des populations [Kimura, 1983]. La probabilité de fixation dépend de la taille de la population effective [Halpern and Bruno, 1998, Yang and Nielsen, 2008] et de la fitness relative entre les codons. Si l'on note w^i_a la fitness de l'acide aminé a en position i de la séquence protéique s, et $f^i_a = \ln w^i_a$, alors la probabilité de fixation dépend de :

$$\Delta f^i = f^i_{a'} - f^i_a, \quad (1.15)$$

qui est la différence de fitness entre les deux acides aminés a et a' au site i. Si $\Delta f^i > 0$, alors cela signifie que a' est plus adapté que a ($\Delta f^i \ll 1$). La probabilité de fixation, pour un site i donné, devient [Halpern and Bruno, 1998] :

$$p^{fix}_i(aa') = p^{fix}_i(\Delta f^i, N) = \frac{1 - e^{-2\Delta f^i}}{1 - e^{-4N\Delta f^i}} \simeq \frac{2\Delta f^i}{1 - e^{-4N\Delta f^i}}. \quad (1.16)$$

Si l'on note $F^i = 4Nf^i$, on peut donc écrire que :

$$2N \cdot p^{fix}_i(aa') = \frac{\Delta F^i}{1 - e^{-\Delta F^i}}. \quad (1.17)$$

A chaque site, est associé un vecteur de 20 coefficients de fitness, $(F^i_a)_{a=1..20}$, rendant donc la sélection spécifique à chaque site. Traiter les sites indépendamment de cette manière revient à poser que les fitness sont multiplicatives [Halpern and Bruno, 1998], c'est à dire que la fitness d'une séquence s, w_s, est exprimée de la manière suivante :

$$w_s = \prod_i w^i_{s_i}, \quad (1.18)$$

où s_i représente l'acide aminé au site i de la séquence s. Ainsi, on peut poser que la fonction de sélection au niveau de l'ensemble de sa séquence, F, est additive, de la forme :

$$F(s) = \sum_i F^i(s_i), \quad (1.19)$$

où i court sur l'ensemble des sites de la séquence. L'avantage de cette hypothèse est qu'elle permet de factoriser le processus de substitution en processus indépendants (61×61) en chaque position.

La sélection est ici purificatrice, et représente donc une tentative d'implémentation de la théorie neutre de Kimura, ou plus exactement de la théorie quasi-neutre de Kimura et Ohta, puisqu'il reste possible d'accepter des mutations légèrement délétères.

Avec leur modèle, Halpern et Bruno montrèrent l'importance de modèles de type mutation/sélection. En comparant des séquences codantes très divergentes, et alors que les autres modèles étudiés sous-estimaient les longueurs de branches (et donc considéraient ces séquences comme plus proches qu'elles ne n'étaient réellement), leur modèle estimait correctement ces distances évolutives.

Le principal intérêt d'un tel modèle mutation/sélection, est d'essayer de séparer les informations provenant uniquement du processus mutationnel (au niveau des séquences nucléotidiques) de la sélection (opérant au niveau des séquences protéiques) afin de mieux comprendre les contributions relatives de chacun à l'évolution du gène, à la fois sur des petites distances évolutives (par exemple pour une unique substitution) mais également sur des plus grandes distances évolutives (le long d'un l'arbre phylogénétique par exemple).

Cependant, de même que cela a été reproché aux modèles à acides aminés, les coefficients de sélection associés aux vingt acides aminés sont déterminés de manière empirique dans le modèle de Halpern et Bruno. Et, si ces coefficients de sélection représentent une sélection explicite, cette sélection n'est pas plus expliquée que dans les modèles précédents. Mais la formulation séparée des termes de mutation et de sélection est extrêmement intéressante d'un point de vue mécanistique : il ne reste plus, dans le modèle décrit précédemment, qu'à remplacer la sélection empirique par une nouvelle fonction de sélection plus explicite. Dans notre cas, nous avons choisi une fonction de sélection reliant la séquence d'une protéine à sa structure.

1.3.2 Modèle d'évolution soumis à des contraintes structurales

> When we consider the action of natural selection at the molecular level, we must keep in mind that higher order (i.e., secondary, tertiary and quaternary) structures rather than the primary structure (i.e., amino acid sequence) are subject to selective constraint, usually in the form of negative selection, that is, the elimination of functionnaly deleterious mutant.
>
> *M. Kimura & T Ohta* (1974) On some principles governing molecular evolution.

Un modèle mutation/sélection particulièrement intéressant serait celui qui paramètrerait la sélection d'après la fonction de la protéine codée. Cependant, bien que les fonctions de certaines protéines soient extrêmement bien connues, il semble actuellement impossible de créer un modèle général où la pression de sélection serait immédiatement dépendante de la fonction de la protéine. D'un autre côté, la fonction de la protéine est intrinsèquement liée aux différents niveaux de sa structure (secondaire, tertiaire, et quaternaire). Si l'on prend l'exemple de la drépanocytose [18], la mutation d'un acide aminé dans le gène codant pour l'hémoglobine conduit, dans des conditions d'oxygénation faibles, à la formation de fibres d'hémoglobine qui déforment complètement les globules rouges et qui les rendent beaucoup moins performants pour le transport d'oxygène. Même s'il s'agit ici d'une liaison entre les molécules d'hémoglobine qui empêche le transport optimal de l'oxygène, cet exemple permet de montrer une dépendance entre la fonction et la structure d'une protéine. D'autres protéines montrent également une dépendance entre leur fonction et leur structure, comme par exemple les enzymes, dont le repliement dans l'espace conditionne la formation du site actif, mais également la liaison au ligand et sa libération après la réaction. Ainsi, introduire une contrainte entre la structure et la séquence d'une protéine, au sein de la probabilité de fixation dans un modèle SC, semble une piste intéressante. De plus, modéliser l'écart à l'homogénéité substitutionnelle entre les sites, via la contrainte imposée par la structure tridimensionnelle, permet d'inclure à l'intérieur d'un modèle une dépendance entre les différents sites de la séquence nucléotidique, et ainsi contourner le problème lié à la traditionnelle hypothèse de sites évoluant de manière indépendante [Robinson et al., 2003, Choi et al., 2007].

La structure d'une protéine est liée à sa séquence en acides aminés. Par exemple, les acides aminés hydrophobes ont tendance à se regrouper les uns avec les autres pour former le cœur de la protéine, alors que les acides aminés hydrophiles ont tendance à s'étaler sur la périphérie de la protéine. Bien entendu, les relations entre la séquence et sa structure ne sont pas exactement définies d'une manière transposable analytiquement, et l'intervention de l'environnement (e.g. les protéines chaperonnes) joue également lors du repliement de la protéine. Cependant, ainsi que nous le verrons dans le chapitre 2, il existe déjà plusieurs fonctions essayant de définir les liens unissant la séquence protéique à sa structure.

En 1999, Bastolla et al. proposèrent un modèle d'évolution neutre, conditionnant l'évolution de la séquence d'une protéine à sa structure [Bastolla et al., 1999]. Les mutations neutres étaient les seules autorisées et, pour qu'une mutation soit acceptée, la nouvelle séquence devait vérifier les propriétés suivantes : un repliement rapide dans la bonne

18. également appelée anémie falciforme.

structure, laquelle devait être stable thermodynamiquement. Afin de modéliser la dépendance structure/séquence, ils utilisaient une fonction mesurant une distance entre la séquence native et la séquence proposée (cf. 2.2.6.4). Cependant, leur modèle impliquait une structure définie en treillis (cf. 2.1.2.1) et donc était limitée pour la taille de la protéine considérée. De plus, ils ne permettaient, dans ce premier modèle, que les mutations réellement neutres, toutes les autres étant considérées comme létales. Des améliorations de ce modèle ont été proposées [Bastolla et al., 2006] avec des variations dans la description de la fonction reliant la séquence protéique à sa structure.

En 2001, Parisi et Echave proposèrent un modèle d'évolution liant explicitement la structure tridimensionnelle des protéines à leurs séquences, par une fonction de score, S_{dist}, dépendant d'une structure de référence, considérée comme fixe au cours du temps [Parisi and Echave, 2001]. Étant donné une protéine particulière, S_{dist} est une fonction de la différence entre l'énergie de la séquence testée (en forçant son repliement dans la structure de référence) et celle de la séquence de référence repliée dans cette même structure, qui est sa structure native. Une séquence proposée était acceptée si $S_{dist} < S_{div}$, où S_{div} représentait la tolérance de divergence acceptée par le modèle d'évolution. Ainsi, dans le modèle proposé par Parisi et Echave, on peut décrire le taux de substitution entre c et c' :

$$R_{cc'} = Q^{mut}_{cc'} \cdot 2N \cdot p^{fix}(aa'), \tag{1.20}$$

où

$$p^{fix}(aa') = \begin{cases} \frac{1}{2N} & \text{si } S_{dist} < S_{div} \\ 0 & \text{sinon} \end{cases}, \tag{1.21}$$

formant ainsi la définition d'un modèle purement neutre. La méthode semblait donner de bons résultats, malgré quelques limites : une probabilité de choix peu flexible (la mutation était acceptée avec une probabilité égale à $1/2N$ si $S_{dist} < S_{div}$ et rejetée dans tous les autres cas) et l'utilisation du modèle de mutation de Jukes Cantor, présenté à l'eq. (1.10).

A partir de ce modèle, Robinson et al. proposèrent un modèle d'évolution soumis à des contraintes structurales plus modulable [Robinson et al., 2003]. Soient σ et σ' deux séquences nucléotidiques plus proches voisines, où le nucléotide b dans le codon c a été remplacé par b' dans c'. Soient s et s' les séquences protéiques codées respectivement par σ et σ', telles que l'acide aminé a dans s a été remplacé par a' dans s'. Soit γ la structure protéique associée aux deux séquences s et s'. Alors, le taux de substitution entre c et c' peut alors être défini ainsi :

$$R_{cc'}^{sub} \propto \begin{cases} \pi'_b & \text{si la mutation est une transversion synonyme} \\ \kappa\pi'_b & \text{si la mutation est une transition synonyme} \\ \pi'_b\omega e^{H(s|\gamma)-H(s'|\gamma)} & \text{si la mutation est une transversion non synonyme} \\ \kappa\pi'_b\omega e^{H(s|\gamma)-H(s'|\gamma)} & \text{si la mutation est une transition non synonyme} \end{cases}, \tag{1.22}$$

où κ correspond au ration transition/transversion, et :

$$H(s|\gamma) - H(s'|\gamma) = \alpha \cdot (E_{acc}(s|\gamma) - E_{acc}(s'|\gamma)) + \epsilon \cdot (E_{cont}(s|\gamma) - E_{cont}(s'|\gamma)). \tag{1.23}$$

$E_{acc}(s|\gamma)$ représente l'énergie d'accessibilité au solvant de la séquence s et $E_{cont}(s|\gamma)$ son énergie de contact. α et ϵ sont des coefficients de sélection liés à la structure : ils sont (tous les deux) égaux à zéro si l'évolution de la séquence est indépendante de la structure (et l'on se retrouve alors avec un modèle à codons tel que décrit dans la section 1.2.5), et strictement positifs si l'évolution de la séquence dépend de la structure [19]. Ces paramètres, tout comme le paramètre ω, permettent de moduler l'influence de la fonction de sélection. Le modèle ainsi présenté permet de moduler à sa guise la probabilité de fixation, correspondant ici à $\frac{1}{2N} \cdot e^{H(s|\gamma)-H(s'|\gamma)}$, et donc de la rendre plus flexible que celle présentée par Parisi et Echave. En effet, si une séquence s' a une moins bonne (resp. meilleure) énergie que s, elle ne sera rejetée (resp. acceptée) que proportionnellement à $e^{H(s|\gamma)-H(s'|\gamma)}$.

Cette équation pour la probabilité de fixation, bien qu'intuitivement assez facile à appréhender, n'a pas vraiment de justification en terme de génétique des populations. Il serait en principe préférable d'utiliser l'équation (1.17) ainsi que l'ont fait Bruno et Halpern dans leur modèle site-spécifique (section 1.3.1). En fait, bien que Halpern et Bruno aient développé leur modèle mutation/sélection quelques années auparavant, leur formulation est restée longtemps ignorée dans les travaux sur les modèles soumis à des contraintes structurales, si bien que la plupart des travaux publiés [Robinson et al., 2003, Rodrigue et al., 2005], ainsi que ceux qui sont présentés dans cette thèse, se sont fait sur la base du modèle de l'équation (1.22). Ce n'est que récemment, sous l'impulsion de Jeffrey Thorne que les modèles ont été reformulés pour être plus en conformité avec la génétique des populations [Thorne et al., 2007].

Contrairement au modèle présenté dans [Bastolla et al., 1999], les mutations acceptées ne sont pas forcément purement neutres, et il s'agit donc ici d'un modèle quasi-neutre. Cependant, ce modèle prometteur n'était applicable en l'état qu'à des paires de séquences codantes, à cause de la méthode d'échantillonnage MCMC sous-jacente.

19. Il est techniquement possible que α ou ϵ soient inférieurs à zéro, mais ils n'ont alors aucune signification biologique.

1.3.3 Implémentation d'un modèle SC

C'est à partir des idées proposées par Robinson et al. qu'a été développé le modèle proposé par Rodrigue et al, qui constitue le cadre méthodologique dans lequel s'articule cette thèse. En reprenant une formulation de la probabilité de fixation basée sur celle de Robinson et al., ce modèle [Rodrigue et al., 2005] généralisait à plus de deux taxons la technique d'échantillonnage proposée par Robinson et al. A partir d'une topologie, d'un alignement de séquences et d'un jeu de paramètres appliqués à la relation structure/séquence, le modèle vise non pas à fournir la meilleure phylogénie, mais à comprendre les relations liant les séquences (nucléotidique et protéique) et la structure tridimensionnelle de la protéine, au sein de l'évolution. Toutefois, ce modèle a d'abord été formulé au niveau des acides aminés, et ce modèle étant par nature un peu trop phénoménologique (les mutations nucléotidiques synonymes ne sont par exemple pas prises en compte), le modèle a par la suite été reformulé en tant que modèle à codons, séparant ainsi la contribution de la mutation, appliquée sur la séquence nucléotidique, et la contribution de la sélection, appliquée sur la séquence protéique.

Différents modèles de mutation ont été intégrés dans ce programme (dont notamment $GY-F61$), mais pour la suite de cette thèse, nous ne considèrerons que le modèle $MG-F1\times4$ qui propose différents avantages. D'abord, ce modèle ne suppose pas des probabilités stationnaires différentes selon le site dans le codon. Il peut en effet sembler étrange de considérer que les différences de distribution des nucléotides entre les différentes positions du codon soient liées à un modèle de mutation qui défavoriserait certains nucléotides à chaque position du codon. Au contraire, il est probablement plus logique, biologiquement parlant, de les considérer comme une conséquence de la sélection au niveau acide aminé. Mais surtout, ce modèle permet de tester la part de la sélection imputable à la relation structure/séquence de la protéine sans aucune autre information liée à la sélection. Ainsi, le modèle mutation/sélection lié à des contraintes structurales présenté ici, permet de tester différents modèles de sélection sans redondance (théoriquement) entre les termes de mutation et de sélection.

On peut ainsi définir le taux de substitution de notre modèle :

$$R_{cc'} = \begin{cases} Q^{mut}_{bb'} \cdot e^{H(s|\gamma)-H(s'|\gamma)} & \text{si } a' \neq a \\ Q^{mut}_{bb'} & \text{si } a' = a \\ 0 & \text{si } c \text{ et } c' \text{ ne sont pas plus proches voisins} \end{cases} , \qquad (1.24)$$

où a (resp. a') est l'acide aminé codé par c (resp. c'), $H(s|\gamma) - H(s'|\gamma)$ est une fonction de sélection (qui dépend de la structure et de la séquence de la protéine étudiée) et $Q^{mut}_{bb'}$

est le taux de mutation définie par le modèle $MG - F1 \times 4$ entre deux nucléotides b et b' :

$$Q_{bb'}^{mut} = \pi_{b'} \rho_{bb'}. \tag{1.25}$$

Le modèle mutationnel sous-jacent correspond au modèle GTR [Felsenstein, 1981] à l'exception du traitement des codons stops, qui sont ici soumis à une contrainte purificatrice maximale. De même que dans le modèle présenté dans [Muse and Gaut, 1994], les codons stops sont exclus d'emblée de l'alphabet mutationnel.

Le facteur de sélection $e^{H(s|\gamma)-H(s'|\gamma)}$ doit pouvoir être calculé à chaque fois qu'une mutation est proposée le long de la phylogénie. Plus encore, à chaque étape, il faut calculer le score de tous les séquences plus proches voisines de la séquence en train d'être mutée. C'est à dire qu'il faut calculer $H(s|\gamma) - H(s'|\gamma)$ pour tous les changements possibles de nucléotide σ_j vers un nucléotide d, pour tous les sites $j \neq i$. De ce fait, la fonction $H(s|\gamma)$ doit pouvoir être calculée rapidement, sans demander beaucoup de puissance computationnelle (que ce soit en temps de calcul ou en taille mémoire), car le modèle en lui-même est déjà extrêmement complexe. On s'attachera donc à trouver une bonne fonction $H(s|\gamma)$ qui constitue un arbitrage intéressant entre la complexité et la rapidité de calcul.

Tout cela s'articule au sein d'un modèle d'évolution probabiliste Bayésien. Cela nous permet notamment de définir un modèle entièrement probabiliste, ainsi que d'utiliser des méthodes de Monte Carlo sophistiquées. De plus, il permet d'utiliser deux méthodes différentes de comparaison de modèles. La première méthode est le rééchantillonnage de "données" d'après les paramètres estimés a posteriori (*posterior predictive*), afin de vérifier qu'elles sont semblables aux données originales. La deuxième méthode, à laquelle nous nous attacherons, s'appuie sur le calcul du facteur de Bayes. Il permet de mesurer le fit relatif d'un modèle par rapport à un autre. Le facteur de Bayes sera décrit, avec les autres méthodes numériques probabilistes faisant partie du cadre méthodologique, dans le chapitre 3.

Chapitre 2

Modéliser la sélection dépendant de la structure - Potentiels statistiques

2.1 Introduction

Les modèles d'évolution moléculaire de type mutation/sélection imposent des contraintes précises sur la probabilité de fixation utilisée dans le modèle : une charge computationnelle faible [20], pour une meilleure fiabilité possible. Le modèle repose certes aussi sur une bonne définition du modèle de mutation, mais il dépend principalement de la définition de la fonction de score utilisée pour calculer la probabilité de fixation. Dans le modèle mutation/sélection qui nous intéresse [Rodrigue et al., 2005, Rodrigue et al., 2006, Rodrigue, 2007], cette probabilité de fixation est exprimée à l'aide d'une fonction représentant la dépendance de la séquence s à sa structure tridimensionnelle c, $H(s|c)$. Une manière intuitive de représenter cette fonction $H(s|c)$ est d'utiliser des fonctions d'énergies basées sur les principes de la thermodynamique.

Il existe déjà bon nombre de fonctions permettant de calculer une telle fonction d'énergie, selon deux approches différentes : on peut choisir de se tourner vers des fonctions d'énergies semi-empiriques, qui décrivent jusqu'à un niveau atomique les relations entre les différents atomes de la protéine, à l'aide des propriétés physiques des atomes, ou bien choisir une fonction d'énergie empirique, dont les paramètres sont appris sur des bases de données biologiques : les potentiels statistiques. Ceux-ci sont extrêmement intéressant dans notre contexte, comme nous le verrons un peu plus tard. Ce chapitre se consacre à l'étude de quelques fonctions d'énergies et de paramètres à prendre en compte lors de la

20. 'A chaque substitution proposée, il faut calculer cette probabilité de fixation pour toutes les séquences s' plus proches voisines de s.

création de la fonction d'énergie modélisant la relation entre la séquence et sa structure, mais aussi aux méthodes d'optimisation de potentiels statistiques les plus intéressantes.

2.1.1 Principes généraux

Les travaux préliminaires d'Anfinsen, posent les bases des relations entre les séquences protéiques et leurs structures tridimensionnelles : les protéines naturelles adoptent la structure d'énergie minimale [Anfinsen, 1973]. À partir d'une séquence donnée, il est donc possible de retrouver sa structure en cherchant le couple structure/séquence d'énergie minimale, parmi l'ensemble des structures possibles pour cette séquence. Bien que ce postulat (la structure native est celle d'énergie minimale pour une séquence donnée) soit sujet à controverse [Sohl et al., 1998], il reste le plus usité dans la détermination de fonction d'énergies, dans sa forme initiale ou modifiée : on considère alors que le couple structure/séquence natif est parmi les couples structure/séquence les plus stables thermodynamiquement.

Les lois de la thermodynamique, et plus particulièrement la loi de Boltzmann, rajoutèrent un aspect stochastique lié à la température finie, en supposant qu'une séquence s visite une structure c avec une probabilité $p(c|s)$ définie par :

$$p(c|s) = \frac{e^{-E(s,c)/kT}}{\sum_{c' \in \mathbb{C}} e^{-E(s,c')/kT}} = \frac{e^{-E(s,c)/kT}}{Z_s}, \tag{2.1}$$

où k est la constante de Boltzmann, T la température absolue du système et $\mathbb{C}$ est l'ensemble des structures possibles pour une séquence de la taille de s. Trouver la conformation d'énergie minimale pour une séquence donnée correspond donc à maximiser la probabilité $p(c|s)$.

Une séquence repliée dans différentes structures est représentée dans la figure 2.1. A gauche, le paysage énergétique créé par le repliement de cette séquence dans différentes structures [Goldstein et al., 1992, Finkelstein, 1997], et à droite l'évolution de l'énergie de la protéine lors du repliement [Branden and Tooze, 1999]. La ligne en pointillés bleus représente l'énergie de la séquence non repliée. On observe qu'on peut séparer l'espace des structures pour cette séquence en trois parties. D'abord, il existe des structures inadaptées, c_m, peu nombreuses, dans lesquelles la séquence est mal repliée, et donc les couples (s, c_m) présentent une énergie élevée. Ensuite, il existe des structures compétitives c_c, relativement proches de la structure native, pour lesquelles les couples c, c_c présentent une énergie basse, proches de l'énergie de la séquence repliée dans sa structure native $(\tilde{c})$. Enfin, la troisième catégorie de structures, c_s contient un très grand nombre de structures, dans lesquelles la séquence se replie avec une énergie intermédiaire. Les structures de cette catégorie étant

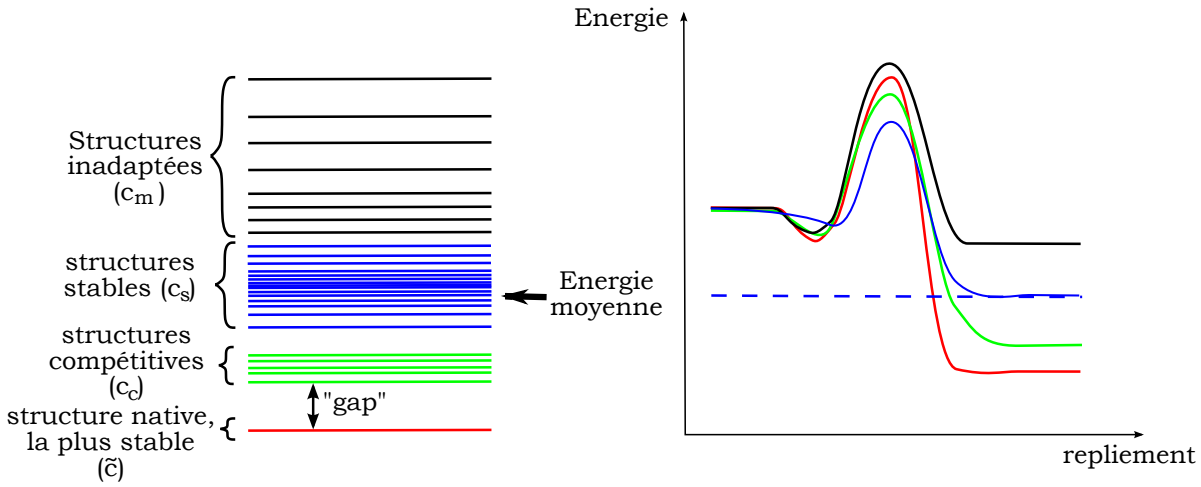


FIGURE 2.1 – Paysage énergétique d'une séquence repliée dans différentes structures.

les plus nombreuses, les couples (s, c_c) correspondant ont potentiellement une grande influence dans la détermination du facteur de normalisation Z_s (eq. (2.1)). D'un autre côté, les structures compétitives, bien que moins nombreuses, ont un poids de Boltzmann beaucoup plus élevé. La question de la contribution relative des structures compétitives proches de la soupe de toutes les autres structures, dans le facteur Z_s, est au centre de la question des méthodes de contraste s'appuyant sur les banques de structures leurres (chapitre 7).

Le problème du *protein folding* consiste à retrouver la structure (soit le repliement soit la conformation) d'une séquence naturelle. Afin de résoudre ce problème, il faut non seulement définir la forme de la fonction d'énergie permettant de hiérarchiser les structures, mais également définir l'espace dans lequel la séquence va se replier. L'ensemble de toutes les structures possibles pour une séquence étant d'une taille bien trop conséquente, on fait généralement appel à des approximations.

2.1.2 Espace des structures

L'espace des structures peut être décrit de différentes manières. La première est de discrétiser l'espace et de créer des modèles de treillis, notamment utilisés dans le cadre de démonstrations théoriques. Une autre méthode de description de structures peut être de créer un ensemble de structures réelles, censé représenter l'ensemble total des structures. Cet ensemble peut représenter des structures extrêmement détaillées ("tous-atomes") ou bien représenter simplement le repliement de la structure, selon l'utilisation à laquelle est voué cet ensemble de structures. Par la suite, nous considèrerons que l'espace des protéines

peut être décrit de manière discrète (soit par les modèles de treillis, soit par les ensembles de structures)[21].

2.1.2.1 Modèles de treillis

Historiquement, la première méthode de description de l'espace des structures consiste à créer un modèle de treillis (*lattice model*). Il s'agit de transformer l'espace réel, cartésien, en un espace discret à deux ou trois dimensions : il existe des treillis carrés [Thomas and Dill, 1996a], cubiques [Yue et al., 1995] mais aussi hexagonaux [Gibbons et al., 2004]. Bien que ce soit une méthode qui simplifie énormément l'espace des structures, il permet de décrire d'une manière exhaustive toutes les structures possibles pour une même séquence. Pour illustrer les méthodes de treillis, j'ai représenté dans la figure 2.2 un treillis cubique de taille 4^3 dans lequel est monté une séquence hypothétique de taille $n = 23$.

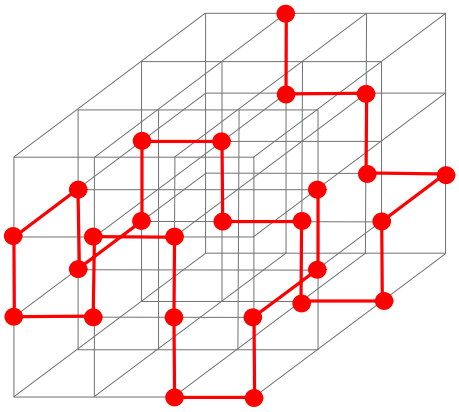


FIGURE 2.2 – Exemple d'un treillis cubique (4^3) sur lequel est montée une séquence hypothétique de taille $n = 23$.

Comme il s'agit d'un espace discret, il est possible de faire l'inventaire exhaustif de toutes les structures possibles (dans ce treillis) de cette séquence de taille 23. Cependant, si l'espace des structures $\mathbb{C}$ est dénombrable dans un tel système, il devient difficilement calculable lorsqu'on augmente la taille de la séquence. En ce qui concerne l'exemple présenté dans la figure 2.2, le véritable ensemble des structures possibles pour cette séquence ne se réduit pas à un espace de taille 4^3 mais demande un espace de description bien plus grand (inclus dans l'espace de taille 23^3). En pratique, on peut réduire l'espace des

21. Il existe une dualité dans la manière de décrire l'espace des structures, car celui-ci peut être considéré d'une manière discrète ou d'une manière continue. Cependant, un espace continu étant inadapté pour notre contexte, nous n'en parlerons pas ici.

structures décrites dans un modèle de treillis en ne visitant que des espaces restreints autour de la conformation native de la protéine [Chiu and Goldstein, 1998b].

De tels modèles de treillis sont cependant problématiques car ils représentent une simplification à l'extrême des structures de protéines. En général, un treillis ne décrit pas des structures réellement existantes, puisqu'il réduit un espace cartésien réel en un espace discret.Cependant, ils sont très utiles dans un cadre théorique, afin par exemple de démontrer la validité d'une approche [Thomas and Dill, 1996b]. Afin de représenter des structures plus réelles que celles créées par les modèles de treillis, une méthode est de faire appel à jeux de structures de protéines réelles.

2.1.2.2 Jeux de structures (*decoys*)

On peut opposer deux buts différents lorsqu'on construit un jeu de structures. D'un côté, on peut chercher à retrouver le repliement d'une protéine, et on cherchera alors à construire un ensemble de repliements qui soit le plus large possible, en utilisant une représentation simplifiée (*coarse-grained*). Ce repliement représente une structure générale, qui peut être utilisé par différentes familles de protéines. D'un autre côté, on peut chercher à retrouver la conformation exacte d'une protéine, jusqu'aux conformations rotamériques des acides aminés et l'on préfèrera alors construire un jeu de structures qui soient proches de la structure cible, en faisant appel à une description extrêmement précise des acides aminés et des atomes constituant la protéine.

Une méthode intuitive pour construire un jeu de structures explorant un large éventail de possibles, est de se baser sur les repliements de protéines existantes. Avec la production intensive de données moléculaires, qui ont eu lieu ces dernières années, on peut supposer que la *Protein Data Bank* (PDB) représente un ensemble exhaustif des repliements possibles. A partir d'une base de données de protéines naturelles, on peut donc construire un jeu de structures représentant la population de toutes les structures possibles (biologiquement), par *threading* [Jones et al., 1992a].

Le *threading* consiste à forcer le repliement de la séquence (ici la séquence native) sur un jeu de structures (généralement issue de la PDB). On peut distinguer deux types de *threading* : le *gapless threading* qui permet de produire des structures sans incohérence au sein de la structure (fig. 2.3 à gauche), et le *gapped threading* qui produit des structures artificielles (et donc permet aussi de produire des structures chimériques) en reliant des parties de protéines entre elles (fig. 2.3 à droite). Pour réaliser des structures leurres sans incohérence dans la structure tridimensionnelle, il faut tout d'abord choisir une protéine p^{thr} dont la séquence est de taille supérieure ou égale à celle de la séquence cible (soit

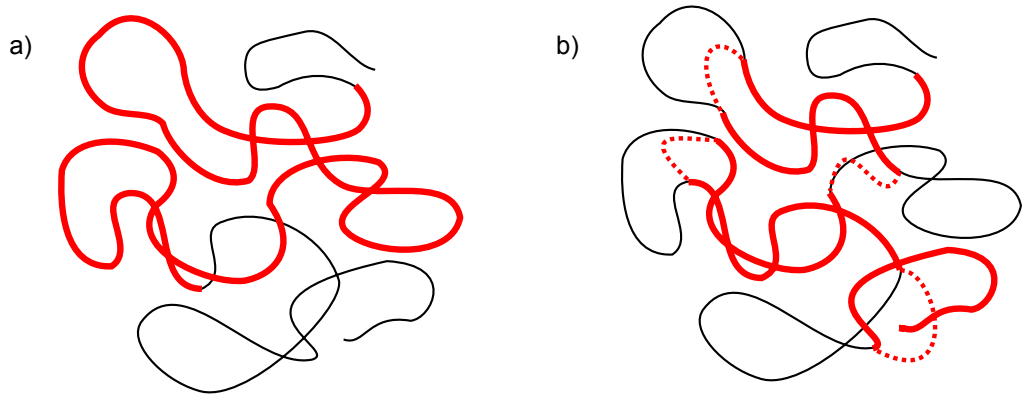


FIGURE 2.3 – *Threading* [Finkelstein et al., 1995a]. a) création d'un couple structure/séquence sans incohérence structurale, b) création d'un couple structure/séquence avec des incohérences structurales, les différentes parties de la nouvelle structure étant reliées par les pointillés rouges.

$n_{thr} \geq n$). Ensuite, il faut choisir aléatoirement une position i au sein de la protéine p^{thr}, telle que $i < n_{thr} - n$. Cette position i et les $n - 1$ suivantes formeront une nouvelle structure dans la banque de structures *threading*.

Afin de réaliser du *gapped threading*, il faut déterminer les différents segments, dans la protéine, qui formeront la nouvelle structure leurre, puis les relier artificiellement. Cette méthode permet de visiter de manière aléatoire tous types de structures et de structures secondaires, alors que la méthode du *gapless threading* permet d'obtenir des structures cohérentes, et surtout des relations entre les différents sites qui sont consistantes avec la biologie. Par exemple, le *gapped threading* pourrait couper la structure à un mauvais endroit, et placer deux structures incompatibles côte à côte. D'un autre côté, ce *gapped threading* permet notamment de faire des déplacements de structures secondaires : on peut par exemple prendre un feuillet β d'une longueur l et l'échanger avec une hélice α de même longueur.

Les jeux de structures créées par *threading* permettent de couvrir un grand nombre de structures possibles, mais également biologiquement plausibles (puisque issues de la PDB). On peut supposer qu'ainsi on peut couvrir quasiment tout l'espace des structures autorisées, ce qui permet d'obtenir l'énergie moyenne de la séquence repliée dans des structures alternatives (cf. fig. 2.1).

D'un autre côté, on peut construire un jeu de structures afin qu'elles soient les plus proches possibles de la structure native. Pour cela, on peut par exemple, à partir de la structure native, utiliser des programmes de dynamique moléculaire, pour déplacer les angles de torsion dans la structure [Rajgaria et al., 2008] et former ainsi une nouvelle

structure leurre. Un ensemble de structures construit de cette manière est typiquement utilisé pour évaluer le *gap* d'énergie présenté dans la figure 2.1. Il faut tout de même noter que même au sein des méthodes qui permettent de générer des jeux de structures proches de la structure native, le *gapped threading* et le *gapless threading* peuvent également être utilisés.

Decoys 'R' Us [Samudrala and Levitt, 2000] est une base de données, regroupant plusieurs ensembles de conformations alternatives et incorrectes de protéines, obtenues de différentes manières. Par exemple, un jeu de structures [Samudrala et al., 1999] a été obtenu en générant, pour une protéine donnée, toutes les structures possibles au sein d'un treillis tétrahédrique. Ces structures furent alors hiérarchisées en fonction d'une fonction de score, afin de ne conserver que les structures les plus proches de la conformation native. Enfin, quelques unes des meilleures structures furent raffinées en utilisant plusieurs fonctions d'énergie différentes. Un autre jeu de structures présenté dans la base de données Decoys 'R' Us fut généré à l'aide d'une procédure de recuit simulé [Simon et al., 1997]. Des protéines ayant des similarités de séquences locales, étaient cassées en plusieurs fragments, lesquels étaient ensuite associés les uns avec les autres à l'aide une fonction de score bayésienne. Ainsi, plusieurs structures leurres proches de la structure native furent générées.

Un jeu de decoys, n'apparaissant pas dans Decoys'R'Us, a été proposé en 2004 puis étendu en 2008, par Floudas et al [Loose et al., 2004,Rajgaria et al., 2008] qui proposaient un potentiel basé sur la génération de structures leurres. La méthode de construction du jeu de structures fonctionne en plusieurs étapes. Pour une protéine donnée, ils déterminent d'abord une séquence secondaire puis un cœur hydrophobe, avant de créer différentes structures en modifiant des angles de torsions à l'aide du programme DYANA [Günter et al., 1997], puis de filtrer les structures afin qu'elles aient une déviation par rapport à la structure native inférieure à une valeur seuil. Ils ont ainsi généré ainsi environ 500 structures leurres pour chacune des 1400 protéines, créant ainsi le plus gros jeu de structures leurres existant actuellement.

2.2 Quelle forme d'énergie ?

Une fois l'espace des structures bien défini, il faut choisir quelle fonction d'énergie utiliser pour déterminer le lien entre la séquence et la structure. Cette fonction d'énergie sera notée $E(s, c)$ par la suite, sachant que $H(s|c)$ (la fonction mesurant le fit entre la séquence et la structure dans le modèle d'évolution) sera une fonction de $E(s, c)$. On peut séparer l'ensemble des fonctions d'énergies en deux types : le premier regroupe les champs

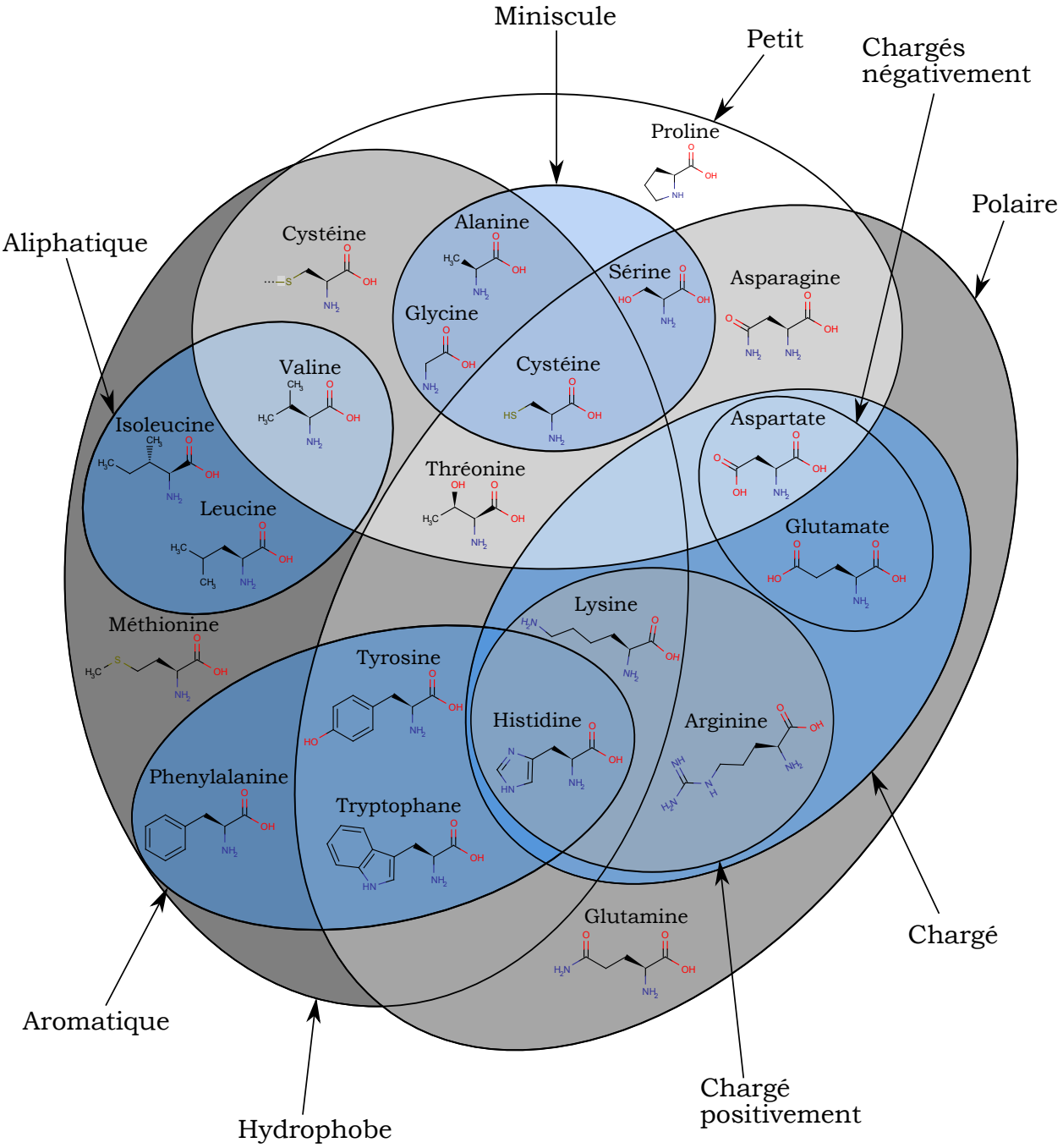


FIGURE 2.4 – Un diagramme de Venn des propriétés des protéines, d'après [Livingstone and Barton, 1993].

de force semi-empiriques, également appelés fonctions d'énergies thermodynamiques, car elles essayent de capter au mieux les relations thermodynamiques entre les atomes ; et le deuxième représente les potentiels statistiques, c'est à dire des paramètres empiriquement appris sur les données. Cependant, avant de détailler quelques unes de ces fonctions d'énergie, je commencerai par présenter rapidement quelques propriétés des acides aminés.

2.2.1 Quelques propriétés

Tous les acides aminés sont composés de la même manière : NH2-CHR-COOH, où R représente le radical, ou la chaîne latérale, de l'acide aminé [22]. Les acides aminés sont reliés entre eux par des liaisons peptidiques, covalentes, et planes : NH2-CHR_1-COO-NH-CHR_2-COOH, où R_1 et R_2 sont les deux radicaux des acides aminés reliés par la liaison peptidique. La succession des liaisons peptidiques forme le squelette de la protéine, qui détermine donc le repliement de la protéine. Les chaînes latérales, différentes selon chaque acide aminé, ont plusieurs conformations dans l'espace, puisque les atomes sont partiellement libres d'effectuer des rotations autours de leurs liaisons covalentes simples. Chaque conformation différente de l'acide aminé est appelée un *rotamère*. Les rotations effectuées par les atomes autour des liaisons covalentes ne sont cependant pas entièrement libres, et certaines conformations sont interdites, alors que d'autres sont très présentes dans les protéines. Afin de recenser les rotamères possibles pour les acides aminés, il existe des banques de rotamères pour chaque acide aminé (voir par exemple [Dunbrack and Karplus, 1993]), regroupant des ensembles plus ou moins exhaustifs de conformations, où les conformations sont hiérarchisées en fonction de leur abondance.

La chaîne latérale, qui induit toute la spécificité des acides aminés, inclut des atomes d'hydrogène, de carbone, d'oxygène, d'azote ou de souffre selon des formes très variées, comme le montre les formules chimiques représentées dans la figure 2.4. Cette figure représente aussi une classification possible des acides aminés selon leurs propriétés chimiques, d'après [Livingstone and Barton, 1993]. On y a représenté deux fois la cystéine, selon si elle établit un pont disulfure, covalent, avec une autre cystéine (ce qui stabilise la structure tertiaire de la protéine), ou si elle reste seule. On notera aussi que la chaîne latérale de la proline se lie de manière covalente au squelette de la protéine, créant ainsi des angles très particuliers (comme par exemple une bifurcation dans les feuillets β) dans le repliement de la protéine. Certains acides aminés sont très petits, comme la glycine, dont la chaîne latérale ne comporte qu'un hydrogène. D'autres acides aminés aromatiques sont bien plus

22. Il existe deux isomères possibles pour chaque acide aminé, mais les L-énantiomères semblent être généralement utilisés dans les protéines naturelles.

gros, comme la phénylalanine, ou presque plats, comme le tryptophane. Certains sont très flexibles, d'autres sont extrêmement rigides. Les acides aminés hydrophobes ont tendance à se regrouper entre eux, à l'intérieur du cœur de la protéine, alors que les acides aminés hydrophiles sont plus souvent retrouvés à la surface de la protéine. Les acides aminés de même charge se repoussent, et attirent la charge opposée...

Toutes ces propriétés, et le choix, par l'évolution, de la position de chacun d'eux dans une séquence permettent le repliement tridimensionnel et l'accomplissement de la fonction de la protéine. Il serait possible de décrire précisément toutes les caractéristiques des acides aminés, et comment ils sont impliqués dans les différentes structures (secondaire, tertiaire, quaternaire) de la protéine. Cependant, les quelques propriétés évoquées plus haut sont suffisantes pour appréhender sommairement le contexte biologique lié à ces propriétés physico-chimiques.

2.2.2 Champs de force semi-empiriques

Les différentes propriétés des acides aminés, évoquées dans la section précédente, permettent d'aider à la détermination des fonctions d'énergies et notamment des champs de force semi-empiriques (appelés ainsi car les paramètres ont initialement été appris sur des bases de données biologiques). Ces champs de force simulent des interactions physiques, thermodynamiques, et représentent actuellement les fonctions d'énergies les plus mécanistiques (dans le sens où ils expriment une énergie dépendant des principes fondamentaux de la physique) que l'on peut trouver. Pour utiliser de telles fonctions d'énergie, il est nécessaire de développer la structure jusqu'au niveau atomique (soit tous les atomes, soit uniquement les atomes lourds), et d'exprimer les relations entre les atomes de manière explicites. L'équation (2.2) représente l'équation d'un des champs de force les plus usités, AMBER [Case et al., 2008] :

$$E_{\text{total}} = \underbrace{\sum_{\text{bonds}} K_r(r - r_{eq})^2}_{1} + \underbrace{\sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2}_{2} + \underbrace{\sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)]}_{3} + \underbrace{\sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]}_{4} + \underbrace{\sum_{\text{H-bonds}} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right]}_{5} \tag{2.2}$$

Le premier terme représente l'énergie lié aux liaisons covalentes entre les atomes, où r représente la distance séparant deux atomes et r_{eq} est la distance d'équilibre qui maximise la liaison covalente entre deux atomes. Cette énergie devient rapidement nulle quand

$r > r_{eq}$. Le deuxième terme représente l'énergie due aux électrons orbitaux impliqués dans la liaison covalente (les deux orbites atomiques se fondent en une seule orbite autour des deux atomes et de la liaison). Le troisième terme représente l'énergie liées aux torsions des liaisons entre les atomes (torsions générées par la formation des doubles liaisons). Les autres termes représentent l'énergie impliquant des atomes non liés par des liaisons covalentes. Le quatrième terme représente les interactions de van der Waals (un terme de répulsion fort à courte distance, et un terme d'attraction à moyenne distance), et les interactions électrostatiques (les atomes ayant des charges inversées s'attirent). Le dernier terme représente l'énergie stabilisatrice des liaisons hydrogènes entre les molécules (les deux molécules reliées se partagent l'hydrogène impliqué dans la liaison).

Il existe d'autres champs de force semi-empiriques : CHARMM [MacKerel Jr et al., 1998], GROMOS [Scott et al., 1997], ECEPP [Momany et al., 1975],... qui utilisent des définitions semblables pour décrire les énergies d'interaction. Nous ne les détaillerons pas ici, mais elles impliquent toujours des termes faisant intervenir les liaisons covalentes et d'autres faisant intervenir les liaisons non covalentes (énergie de Van der Waals, électrostatique, ou causée par la formation des liaisons hydrogènes).

Ces champs de force semi-empiriques sont souvent testés lors du CASP, Critical Assessment of techniques for protein Structure Prediction : à partir de séquences protéiques (dont la structure cristalline a été reconstruite mais non encore publiée), les fonctions doivent retrouver les structures correspondantes. Un classement est ensuite effectué entre les différentes méthodes, afin de déterminer lesquelles ont fait les prédictions les plus conformes à la structure dévoilée. Un excellent test pour ces fonctions qui se révèlent de plus en plus efficaces.

Cependant, si tous ces champs de forces semi-empiriques sont de plus en plus complexes et précis, ils sont généralement coûteux en temps de calcul, et l'on ne peut utiliser de telles fonctions pour modéliser les relations entre la structure et la séquence à l'intérieur d'un modèle génétique déjà complexe. Une alternative, plus empirique mais également plus rapide à calculer, peut être envisagée : il s'agit des potentiels statistiques.

2.2.3 Potentiels statistiques

L'avantage computationnel offert par les potentiels statistiques repose sur la représentation simplifiée (*coarse-grained*) choisie. L'une des plus simples, à laquelle nous nous attacherons, représente l'acide aminé comme un unique corps (ou pseudo-atome), formant alors une structure en "collier de perles". Cependant, il existe bien d'autres représentations, plus ou moins simplifiées : il existe par exemple des représentations à deux corps

(C_α et centre de masse de la chaîne latérale l'acide aminé) *coarse-grained*, trois corps (C_α et deux pseudo-atomes pour représenter la chaîne latérale), ou même quatre corps (représentation des atomes N, C_α, C et un pseudo atome pour la chaîne latérale)[23]. D'un autre côté, on peut également réduire la complexité du modèle en réduisant le nombre de classes d'acides aminés. Par exemple, le modèle le plus simplifié, lié aux *go-like models* (également appelés modèles HP), sépare les acides aminés en deux classes différentes, H et P, hydrophiles et hydrophobes.

Un deuxième avantage vient de la manière de construire les potentiels statistiques. En effet, ceux-ci sont bâtis de manière à mimer la loi de Boltzmann présentée à l'equation (2.1) :

$$p(c|s) = \frac{e^{-E(s,c)/kT}}{\sum_{c' \in \mathbb{C}} e^{-E(s,c')/kT}} = \frac{e^{-E(s,c)/kT}}{Z_s}.$$

Ainsi, le paysage énergétique des structures décrit par les potentiels statistiques est analogue à celui décrit par les champs de force semi-empiriques : pour une séquence donnée, la meilleure structure a l'énergie la plus basse.

Un dernier avantage en faveur des potentiels statistiques est qu'ils sont appris sur des bases de données, ce qui leur donne leur caractère phénoménologique. On peut donc espérer que dans le potentiel choisi, on captera l'essentiel des forces d'interactions définies par les champs de force semi-empiriques, mais dans une forme simplifiée. Ainsi, un potentiel de contact entre deux acides aminés non reliés par une liaison covalente devrait inclure une partie de l'énergie de Van der Waals, mais également une partie de l'énergie de liaison électrostatique et une partie de l'énergie liée aux liaisons hydrogènes.

Dans le cadre des modèles d'évolution soumis à des contraintes structurales, nous avons décidé de nous orienter vers les potentiels statistiques, au vu des trois avantages que je viens d'évoquer. Par la suite, je ne parlerai donc que des potentiels statistiques, en omettant les champs de force semi-empiriques. Dans le cadre d'un modèle mécanistique d'évolution, ces derniers seraient intuitivement les plus adaptés, car ils représentent les relations entre les atomes d'une manière mécanistique. Cependant, les champs de force semi-empiriques seraient trop gourmands en ressources (l'énergie doit être calculée pour chaque proposition de mutation, pour le site considéré et pour les séquences plus proches voisines) pour que le modèle d'évolution reste à un niveau de complexité raisonnable.

23. A noter qu'il existe également des potentiels statistiques atomiques.

2.2.4 Formes générales de potentiels statistiques

Les champs de force semi-empiriques (issus des lois de la thermodynamique) comprennent beaucoup de paramètres qui ne sont pas facilement formalisables dans une représentation simplifiée utilisant des potentiels statistiques. Par exemple, les interactions liées aux liaisons hydrogènes ne peuvent être décrites de la même manière dans un potentiel statistique. Cependant, on peut imaginer qu'une part de ces interactions soit captée par un simple potentiel de contact entre acides aminés. Comme on vient de l'évoquer, il existe différentes manières de formaliser la représentation des acides aminés, qui, chacune, affecte différemment la forme du potentiel statistique associé. De plus, la représentation choisie et la forme du potentiel statistique sont conditionnées par l'utilisation à laquelle est voué le potentiel statistique. Par exemple, si l'on cherche à retrouver les conformations rotamériques d'une protéine, voire même retrouver certaines structures, l'on préfèrera une représentation complexe [Maupetit et al., 2007]. Dans d'autres contextes, et notamment celui des modèles d'évolution SC, où l'on s'intéresse aux acides aminés et non aux atomes qui les composent, on préfèrera une structure *coarse-grained*.

A cela s'ajoutent d'autres simplifications liées à la structure : on peut par exemple représenter la structure tridimensionnelle par une simple matrice de contact ou bien essayer de représenter les distances entre toutes les paires, triplets ou quadruplets possibles d'atomes.

Le potentiel qui sera utilisé dans ce mémoire est un potentiel statistique extrêmement simple, composé de deux termes, puisque cette thèse se concentre sur des méthodes d'optimisation de ce potentiel et non pas sur les différents termes pouvant être utilisés. La forme du potentiel décrite par la suite présente l'avantage d'être généralisable à bien d'autres formes de potentiel [Kleinman et al., Submitted]. La suite de cette section (2.2.4.1 et 2.2.4.2) se concentre donc sur la manière de décrire ces deux termes : le premier décrivant un potentiel d'interaction entre deux acides aminés, et l'autre représentant l'accessibilité au solvant.

2.2.4.1 Potentiel d'interaction

Le premier terme intéressant que l'on peut développer dans un potentiel statistique est un potentiel d'interaction entre deux acides aminés, ou entre deux pseudo-atomes. Cependant, il existe des potentiels qui prennent en compte des interactions mettant en jeu trois [Rossi et al., 2001], voire quatre corps [Feng et al., 2007]. Ces corps peuvent être des acides aminés, dans le cadre de la représentation la plus simple mais également des atomes ou de pseudo-atomes dans le cadre de représentations plus complexes. En outre,

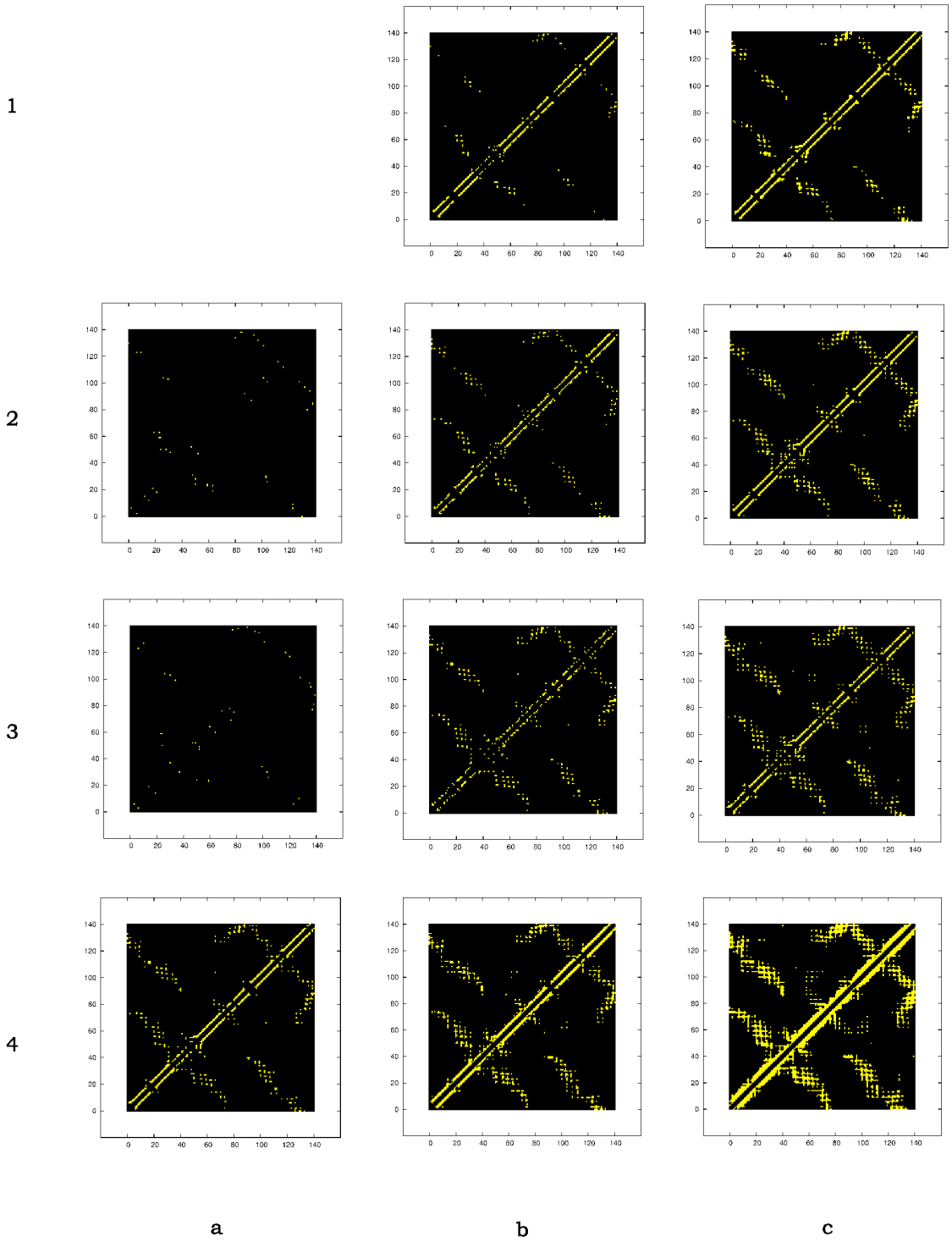


FIGURE 2.5 – matrices de contact obtenues par protInfo (CL Kleinman, communication personnelle), pour la chaîne B de l'hémoglobine de Dasyatis akajei, un poisson cartilagineux. Ligne 1 : le centre de contact est le C_α, ligne 2 : le centre de contact est le C_β, ligne 3 : le centre de contact est le centre de masse de la chaîne latérale, ligne 4 : définition "tous-atomes" des contacts. Colonne a : $v_{seuil} = 4,5$ Å, colonne b : $v_{seuil} = 6,5$ Å, colonne c : $v_{seuil} = 8,5$ Å.

ces interactions peuvent être définies de plusieurs manières : on peut par exemple définir des matrices de contact entre les corps, ou bien utiliser un potentiel de distance, voire même un potentiel de distance discrétisé.

Matrices de contact Le potentiel de référence [Miyazawa and Jernigan, 1985], optimisé dans un contexte de *protein folding*, utilise une représentation de la structure par une matrice de contact. Une matrice de contact est une représentation simplifiée de la structure, sous la forme d'une matrice de 1 et de 0, indiquant si deux positions dans la séquence sont en contact dans la structure. Un contact est défini si la distance entre les deux positions est inférieure à une certaine valeur seuil (v_{seuil}), i.e. :

$$C_{ij} = \begin{cases} 1 & \text{si } d_{ij} \leq v_{seuil} \\ 0 & \text{sinon} \end{cases} \tag{2.3}$$

où d_{ij} est la distance séparant deux sites dans l'espace. La forme d'une énergie de contact est alors :

$$E_{contact} = \sum_{1 \leq i < j \leq n} C_{ij} \varepsilon_{s_i s_j}, \tag{2.4}$$

où n est la taille de la protéine. Deux subtilités apparaissent ici : la définition du centre de contact et la distance seuil. Dans la figure 2.5 sont représentées différentes matrices de contact pour une même protéine, en faisant varier les critères de contact et le seuil : 4,5 Å (colonne a), 6,5 Å (colonne b) ou 8,5 Å (colonne c). Notons que pour la protéine représentée ici, il n'y a pas de contact pour un seuil de $4,5$ Å entre les C_α des acides aminés. On peut voir dans la figure 2.5 que la matrice de contact garde globalement la même forme, mais que la structure est plus ou moins bien définie : par exemple pour un contact défini entre les carbones β (C_β), on ne devine pas la structure pour un seuil de $4,5$ Å mais elle apparaît très clairement pour un seuil de $8,5$ Å. On peut aussi voir que la structure secondaire peut également être visible sur une telle représentation simplifiée : les feuillets β parallèles et anti-parallèles se démarquent très nettement (par exemple sur la matrice 2.c).

Dans le potentiel de Miyazawa et Jernigan, la matrice de contact est définie entre les centres de masse des chaînes latérales, pour un seuil de $6,5$ Å. Cependant, deux acides aminés ne sont considérés comme étant en contact qu'à la condition qu'ils soient éloignés de plus de 2 sites dans la séquence : le potentiel étant censé représenter des interactions non covalentes, introduire les interactions entre les acides aminés adjacents causerait une pollution du potentiel par des énergies provenant des liaisons covalentes. L'utilisation du

centre de masse de la chaîne latérale de chaque acide aminé comme le centre des contacts était une nouveauté lorsque Miyazawa et Jernigan présentèrent leur potentiel, car tous les autres potentiels statistiques utilisaient le carbone α (C_α) pour définir les contacts entre deux acides aminés. Cependant, le potentiel représentant des interactions entre les chaînes latérales, il semblait à Miyazawa et Jernigan que le centre de masse de la chaîne latérale était plus approprié pour ce potentiel que n'importe quel atome de la chaîne principale. Actuellement, même si les autres définitions de matrices de contact peuvent encore utiliser les C_α [Furuichi and Koehl, 1998] ou les C_β [Rossi et al., 2001] comme les centres de contact, la définition des contacts la plus usitée reste celle définie par Miyazawa et Jernigan, avec quelques variations pour la distance seuil.

Notons cependant qu'il existe une dernière définition pour les matrices de contact (fig. 2.5 ligne 4) [Mirny and Shakhnovich, 1996, Vendruscolo et al., 2000, Bastolla et al., 2000] : un contact est défini entre deux acides aminés si deux de leurs atomes lourds (c'est à dire n'importe quel atome sauf l'hydrogène) sont distants de moins de v_{seuil} (dans le cas de Bastolla et al, $v_{seuil} = 4,5$ Å). Dans [Vendruscolo et al., 2000], les auteurs comparèrent leur définition de contact, "tous-atomes" à celle considérant le C_α comme centre de contact, en optimisant les potentiels sur le même jeu de données, et en comparant l'aptitude de chacun des potentiels à retrouver le repliement natif parmi un jeu de structures leurres. Ce travail montrait que la définition "tous-atomes" était plus efficace pour retrouver les repliements natifs, mais également que, pour leur critère d'optimisation, le potentiel obtenu à l'aide d'une telle définition de contact avaient un domaine d'apprentissage stable plus large. C'est à dire qu'il était possible de trouver un potentiel qui favorisait toutes les repliements natifs en même temps (domaine d'apprentissage stable), pour un ensemble de protéines plus grand que si l'on utilisait la définition de contact utilisant uniquement les C_α [24]. On peut également noter qu'un tel potentiel prend en compte plus d'interactions qu'un potentiel utilisant une définition des contacts plus simple pour une même valeur v_{seuil} (cf. figure 2.5).

Parmi les potentiels de contact, on peut également citer une définition de contact non binaire, introduisant une gradation des valeurs de la matrice de contact [Rossi et al., 2001]. Au lieu d'assigner une valeur 0 (pas de contact) ou 1 (contact) à leur matrice de contact, C_{ij} était définie de la manière suivante :

$$C_{ij} = \frac{1}{2}\tanh(a_0 - d_{ij}) + \frac{1}{2}, \tag{2.5}$$

où a_0 était une constante arbitrairement fixée à 8 Å. Ainsi, la valeur de C_{ij} est com-

24. Le domaine d'apprentissage devient toujours instable à compter d'un certain seuil, à cause notamment des violations de modèles (cf. 2.2.6.2).

prise entre 0 (pour les sites très éloignés l'un de l'autre) et 1 (pour les sites extrêmement proches), en évitant le système binaire induit par une matrice de contact classique. Malheureusement, ce potentiel de contact ne fut défini qu'en partitionnant l'ensemble des acides aminés en trois classes, revenant ainsi à une représentation un peu plus complexifiée d'un modèle *go-like*.

Distances Les champs de force semi-empiriques expliquent généralement les interactions entre les paires de corps par la formule de Lennard-Jones, le plus souvent par la formulation (6-12) que l'on peut typiquement décrire par :

$$E_{distance}(lm) = \left(\frac{A_{lm}}{r_{lm}}\right)^{12} - \left(\frac{B_{lm}}{r_{lm}}\right)^{6}, \tag{2.6}$$

où r_{lm} est la distance entre les deux corps r et l. Il est à noter que dans la formule initiale de Lennard-Jones, ces deux corps représentaient des atomes et non pas des groupes d'atomes. Évidemment, cette énergie ne peut être qu'approximative lorsqu'on utilise des potentiels statistiques *coarse-grained*. De plus, l'optimisation de tels potentiels se révèle gourmande en puissance et en temps de calcul, sans compter l'apparition de problèmes au voisinage de zéro. En effet, si deux corps sont trop proches, la force de répulsion devient extrêmement élevée, et peut potentiellement causer des instabilités numériques lors de certaines applications. En particulier, j'ai pu le constater lors d'une tentative d'optimisation d'un potentiel Lennard-Jones par une méthode de gradient. De plus, ils seraient probablement trop gourmands au sein d'un modèle SC.

Par contre, on peut définir d'autres potentiels de distance, *coarse-grained*, en utilisant une méthode de distances discrétisées. Cette méthode consiste à partitionner l'espace des distances en plusieurs matrices de contact. Chaque matrice de contact C_{ij}^{m} a alors un seuil différent [Sippl, 1993a, Sippl, 1993b, Kleinman et al., Submitted] et il lui est associé un potentiel de contact qui lui est propre, ε_{ij}^{m}, tel que :

$$E_{distance} = \sum_{1 \leq m \leq M} C_{ij}^{m} \varepsilon_{ij}^{m}, \tag{2.7}$$

où M est le nombre total de classes de distance. Les classes de distance sont construites de manière à ne pas être redondantes : ainsi, si l'on a observé un contact entre i et j pour une classe de distance m, alors cette paire ij ne sera pas considérée en contact dans la classe $m+1$. Le partitionnement de l'espace peut être variable : on peut par exemple déterminer les classes par tranche de $0,5$ ou 1 Å, ou choisir des tailles de classes variables. Une telle définition de potentiel de distances discrétisées est notamment intéressante parce que la complexité du calcul de $E_{distance}$ (le temps de calcul requis pour un potentiel de contact) est le même pour une même valeur v_{seuil}^{max}, quel que soit le nombre de classes considérées.

2.2.4.2 Potentiel d'accessibilité au solvant

Le caractère hydrophobe/hydrophile des acides aminés est un point extrêmement important dans le repliement des protéines. Les acides aminés hydrophobes ont tendance à se regrouper à l'intérieur de la structure, afin de former le cœur de la protéine, alors que les acides aminés hydrophiles ont plutôt tendance à se placer à l'extérieur de la protéine, donc en interaction avec le solvant. Cependant, il existe un certain nombre d'exceptions à cette règle, comme par exemple les protéines trans-membranaires, qui présentent des acides aminés hydrophobes à la surface de leur structure (afin d'être bien intégrés dans la membrane lipidique). D'autre part, certaines protéines intègrent des acides aminés hydrophiles à l'intérieur même de la structure, lesquels sont ensuite stabilisés par des interactions polaires [Shirota et al., 2008]. De plus, d'un point de vue évolutif, il a été observé que l'hydrophobicité semblait être conservée au cours de l'évolution (les acides aminés hydrophobes remplacent d'autres hydrophobes et les hydrophiles remplacent d'autres hydrophiles). Ainsi, une manière de prendre en compte l'hydrophobicité est de définir un potentiel d'accessibilité au solvant.

Cependant, il n'est pas forcément évident de calculer un tel potentiel. L'algorithme de Lee et Richards [Lee and Richards, 1971] permet de déterminer les surfaces accessibles au solvant, mais une fois de plus (comme pour les potentiels de contact), il faut discrétiser l'espace afin de rendre ce paramètre intégrable dans un potentiel statistique [Dehouck et al., 2006].

Sans une définition explicite d'un potentiel d'accessibilité au solvant, les interactions de ce type sont directement intégrés dans le potentiel de contact. Ainsi, le potentiel de Miyazawa et Jernigan inclut indirectement dans le terme de contact un terme dépendant des caractères hydrophiles et hydrophobes des deux acides aminés considérés. La représentation de la structure par un treillis permet de définir la concentration de molécules de solvant autour de chaque acide aminé de la protéine et ils intégraient dans leur potentiel une valeur dépendant de la concentration des molécules de solvant de le voisinage des résidus des structures hypothétiques, à l'aide de l'approximation quasi-chimique (cf. 2.2.5.1).

D'autres formes de potentiels peuvent être proposées, comme celui présenté dans [Kurochkina and Lee, 1995] : le potentiel ("tous-atomes") y est défini comme étant la somme des énergies entre les paires d'acides aminés. Cependant, ce potentiel de contact entre les acides aminés dépend de l'aire enfouie par un couple d'atomes, du rayon supposé d'une molécule d'eau et de la distance entre les deux atomes considérés. Cependant, le potentiel de Kurochkina et Lee présentait une corrélation de $0,96$ avec le potentiel de Miyazawa

et Jernigan, montrant ainsi que leur approximation correspondait au moins en partie à l'approximation quasi-chimique.

2.2.4.3 Autres paramètres

D'autres paramètres peuvent être pris en compte dans les potentiels statistiques. Par exemple, on peut imaginer un potentiel dépendant de la structure secondaire à laquelle les positions appartiennent, ou l'on peut définir un potentiel lié aux angles de torsion (ϕ et ψ) le long du squelette de la protéine. Ainsi, Dehouck et al proposent de présenter un terme du potentiel statistique dépendant de ces angles de torsion, en partitionnant l'espace en sept classes, déterminées par le diagramme de Ramachandran [Dehouck et al., 2006]. Dong et al proposent un potentiel de torsion dépendant du nombre d'observation des angles ϕ et ψ pour chaque site, par rapport au nombre moyen d'observations des angles, en partitionnant l'espace des angles en 36 classes [Dong et al., 2006].

D'autres méthodes pour intégrer les angles de torsion peuvent être proposées, tout comme pour d'autres paramètres comme la structures secondaire. Certains essaient même d'introduire des paramètres de relaxation de la structure : la structure décrite par un fichier PDB n'est généralement qu'une structure parmi toutes celles adoptées par la protéine. Donc, on peut essayer de relaxer un peu la contrainte de la structure native. Cependant, à chaque fois que l'on cherchera à intégrer de nouveaux paramètres dans un potentiel statistique, il sera pratique, voire essentiel, de partitionner l'espace en différentes classes.

2.2.5 Estimer des énergies du potentiel

Un des avantages des potentiels statistiques sur les champs de force semi-empiriques est qu'ils sont entièrement appris sur des bases de données biologiques. En outre, ils sont construits de manière à ce que, pour une séquence donnée, la structure native ait la plus basse énergie par rapport aux autres structures. Il existe différentes méthodes permettant d'estimer de tels potentiels. La première d'entre elles consiste à extraire de manière brute des potentiels à partir des bases de données. Je commencerai d'abord par décrire l'approximation quasi-chimique utilisée par le potentiel de référence [Miyazawa and Jernigan, 1985, Miyazawa and Jernigan, 1996], avant de généraliser la forme de ces potentiels [Sippl, 1990].

2.2.5.1 Approximation quasi-chimique

Si l'idée de créer des potentiels statistiques de contact fut d'abord proposée dans [Tanaka and Scheraga, 1976], le potentiel de référence est celui présenté dans [Miyazawa

and Jernigan, 1985]. A partir d'une base de données de protéines, Miyazawa et Jernigan estimèrent un potentiel de contact entre des paires d'acides aminés (incluant un terme implicite lié au solvant) en utilisant l'approximation quasi-chimique. L'estimation des paramètres du potentiel se faisait par comptage des contacts entre les acides aminés et entre ceux-ci et les molécules de solvant, pour toutes les protéines de la base de données. Afin de modéliser la structure de la protéine et des interactions avec le solvant, ils placèrent chacune des protéines sur un treillis, et chaque emplacement libre restant dans le treillis était considéré comme étant occupé par une molécule de solvant. L'énergie de contact était donc de la forme :

$$E_{contact} = \sum_{1 \leq a \leq 20} \sum_{1 \leq b \leq 20} E_{ab} n_{ab}, \tag{2.8}$$

où n_{ab} est le nombre de contact entre les résidus de types a et b, et E_{ab} est l'énergie physique de contact entre ces deux acides aminés. Afin de pouvoir retrouver celle-ci, on a recours à des estimateurs, e_{ab} et e'_{ab} extraits directement de la base de données. Ces termes sont reliés aux énergies physiques, E, par les relations suivantes :

$$e_{ab} \equiv E_{ab} + E_{00} - E_{a0} - E_{b0} \tag{2.9}$$

$$e'_{ab} \equiv E_{ab} - (E_{aa} + E_{bb})/2, \tag{2.10}$$

où 0 représente une molécule de solvant. Il devient possible de décrire les deux énergies relatives l'une en fonction de l'autre :

$$e_{ab} = e'_{ab} + e'_{00} - e'_{a0} - e'_{b0} \tag{2.11}$$

$$e'_{ab} = e_{ab} - (e_{aa} + e_{bb})/2, \tag{2.12}$$

On peut voir les énergies des équations (2.9) et (2.10) comme les énergies (libres) des réactions chimiques :

$$a{-}0 + b{-}0 \leftrightarrow a{-}b + 0{-}0 \tag{2.13}$$

$$\frac{1}{2}(a{-}a + b{-}b) \leftrightarrow a{-}b \tag{2.14}$$

La constance d'équilibre K de la réaction chimique de l'équation (2.14) est définie comme suit :

$$K = \frac{n_{ab}}{\sqrt{n_{aa} n_{bb}}} = \exp(-e'_{ab}). \tag{2.15}$$

Si l'on considère que l'énergie de la protéine est l'énergie nécessaire pour passer de la protéine dépliée et entièrement solvatée (c_0) à la protéine repliée (c), et que l'on fait l'approximation que l'on peut traiter indépendamment chaque paire $\{n_{ab}\}_{a,b\in\{1..20\}}$ comme des variables aléatoires indépendantes (approximation quasi-chimique), alors :

$$\frac{p(c|s)}{p(c_0|s)} = \exp\left(-\sum_{ab} n_{ab} e_{ab}\right), \tag{2.16}$$

où c_0 est aussi appelé l'état de référence. L'approximation quasi-chimique, qui considère donc que la séquence est en fait une "soupe" d'acides aminés, permet de calculer les énergies qui nous intéressent, e_{ab} et e'_{ab} de la manière suivante :

$$\exp(-e_{ab}) = \frac{\bar{n}_{ab}\bar{n}_{00}}{\bar{n}_{a0}\bar{n}_{b0}} \tag{2.17}$$

$$\exp(-2e'_{ab}) = \frac{\bar{n}^2_{ab}}{\bar{n}_{aa}\bar{n}_{bb}} \tag{2.18}$$

où $\bar{n}_{ab}$ correspond à la moyenne statistique des contacts entre a et b.

Le potentiel de Miyazawa et Jernigan reste encore aujourd'hui la référence [25], et de nombreuses études se sont attaquées à ses faiblesses. Certains ont par exemple éliminé l'hypothèse qui ignore la connectivité de la chaîne principale, d'autres ont montré la grande dépendance de ces types de potentiels envers la taille des séquences natives [Furuichi and Koehl, 1998].

2.2.5.2 Généralisation

On peut noter que, bien que la forme du potentiel soit différente entre celui de Miyazawa et Jernigan et celui présenté dans [Sippl, 1990] (mais également tous les potentiels estimés à partir une base de données [26]), l'idée à la base de l'estimation des énergies reste la même [Thomas and Dill, 1996a, Thomas and Dill, 1996b, Jernigan and Bahar, 1996] :

$$e_{ab} = -\ln\frac{\rho_{ab}}{\rho^*_{ab}}, \tag{2.19}$$

où ρ_{ab} correspond à la probabilité avec laquelle on a observé les résidus a et b en contact et ρ^*_{ab} représente les probabilités (ici fréquences ou nombres de contact) de la configuration homologue (contact, classe de distance, etc...) pour l'état de référence.

25. Dans sa forme première ou redéfinie [Miyazawa and Jernigan, 1996].

26. Le terme *estimé* sera uniquement appliqué à ces potentiels obtenus par comptage, par opposition avec les potentiels optimisés que nous verrons à la section 2.2.6.

Cet état de référence peut être très variable, et sa définition est critique pour tous ces potentiels estimés. Dans le cadre des potentiels du genre de Miyazawa et Jernigan, on peut citer trois types d'état de référence [Jernigan and Bahar, 1996]. Le premier consiste à penser que dans l'état de référence, les acides aminés préfèrent s'apparier entre eux plutôt que de s'associer à des acides aminés de type différent :

$$a{-}a + b{-}b \leftrightarrow 2a{-}b \Rightarrow e^{2e_{ab}} = \frac{n_{ab}^2}{n_{aa}n_{bb}}. \tag{2.20}$$

Le deuxième considère que les acides aminés sont normalement en contact avec le solvant, dans l'état de référence :

$$a{-}0 + b{-}0 \leftrightarrow a{-}b + 0{-}0 \Rightarrow e^{2e_{ab}} = \frac{n_{ab}n_{00}}{n_{a0}n_{b0}}, \tag{2.21}$$

où 0 représente une molécule de solvant [Miyazawa and Jernigan, 1985]. Le dernier état de référence consiste à remplacer la molécule de solvant dans l'équation précédente par un résidu moyen, hypothétique, r :

$$a{-}r + b{-}r \leftrightarrow a{-}b + r{-}r \Rightarrow e^{2e_{ab}} = \frac{n_{ab}n_{rr}}{n_{ar}n_{br}}. \tag{2.22}$$

Dans le cadre du potentiel de Miyazawa et Jernigan, l'état de référence est une combinaison des deux premiers états de référence : d'abord un phénomène de désolvation permet de lier deux acides aminés de même type ensemble, puis un mélange met en contact les deux acides aminés a et b [Thomas and Dill, 1996b].

L'état de référence présenté à l'équation (2.22) a notamment été utilisé par [Betancourt and Thirumalai, 1999] afin de re-estimer les potentiels de Miyazawa et Jernigan et de les comparer à un autre potentiel estimé [Skolnick et al., 1997]. Alors que les deux potentiels originaux présentaient, malgré une corrélation certaine, une dispersion importante, Betancourt et Thirumalai montrèrent notamment qu'en définissant un état de référence différent pour le potentiel de Miyazawa et Jernigan, la dispersion était réduite [27], prouvant ainsi que la définition d'un état de référence est critique pour l'estimation de tels potentiels.

On peut appliquer l'approximation quasi-chimique à des potentiels de distance discrétisés [Sippl, 1990] :

$$\Delta E_k^{ab}(d) = -kT \ln \frac{f_k^{ab}(d)}{f_k(d)}, \tag{2.23}$$

où f_k^{ab} représente la fréquence à laquelle on a observé a et b en contact (avec $f_k^{ab} \neq f_k^{ba}$ pour le potentiel défini par [Sippl, 1990]), à une distance donnée d, et k représente la distance

27. Elle passait de 3,4 à 0,45.

en acides aminés séparant a et b. La fréquence de l'état de référence, $f_k(d)$, représente donc la fréquence à laquelle on observe un contact entre deux acides aminés quelconques séparés par une distance cartésienne d et une distance dans la séquence de k sites.

Parmi les potentiels utilisant l'approximation quasi-chimique, on peut également citer le travail de Buchete et al, qui, en utilisant la formulation de l'équation (2.19), ont créé des potentiels orientés et dépendants de la distance séparant deux acides aminés [Buchete et al., 2004a, Buchete et al., 2004b]. Un potentiel de contact orienté, entre deux acides aminés a et b, dépend des orientations relatives des chaînes latérales de a et b, ce que je traduirais par une dépendance entre les deux rotamères impliqués dans ce contact, et la manière dont ils sont orientés l'un par rapport à l'autre.

Cependant, en 1996, Thomas et Dill montrèrent que les potentiels de contact obtenus à l'aide de l'approximation quasi-chimique (ou des approximations équivalentes) étaient de très pauvres estimateurs statistiques des énergies réelles [28] à cause d'erreurs systématiques, et qu'ils sous-estimaient notamment les énergies défavorables [Thomas and Dill, 1996a, Thomas and Dill, 1996b]. Afin de démontrer que les potentiels estimés n'étaient pas vraiment adaptés, Thomas et Dill construisirent un modèle *exact* de treillis où ils spécifiaient une séquence et toutes les structures possibles dans ce treillis, pour cette séquence, afin de former la base de données d'apprentissage. A partir de ce modèle, ils sélectionnaient une structure native d'après un potentiel (posé comme étant la véritable énergie du système). Ensuite, ils estimèrent les potentiels sur la base de données de structures, et les comparèrent aux potentiels originels. Grâce à leur modèle, ils montrèrent que la connectivité de la chaîne, ignorée par l'approximation quasi-chimique, est importante. En effet, deux acides aminés hydrophiles avec des charges opposées, présents à la surface de la protéine, peuvent être poussés chacun d'un côté de la protéine car le cœur est occupé par des acides aminés hydrophobes. Par l'approximation quasi-chimique, l'énergie d'interaction serait considérée comme non favorable, alors qu'il s'agit en fait d'un phénomène extérieur à cette paire d'acides aminés.

2.2.6 Optimisation directe des énergies

Les méthodes d'optimisation des potentiels sont une alternative à l'utilisation de l'approximation quasi-chimique, en essayant également de correspondre aux données, mais en prenant en compte la connectivité de la protéine.

28. Une conclusion généralement admise, bien qu'elle ait été parfois mise en doute [Mirny and Shakhnovich, 1996].

2.2.6.1 Optimisation itérative

Afin de circonvenir aux problèmes posés précédemment, Thomas et Dill proposèrent une méthode *itérative* pour optimiser des potentiels, ENERGI [Thomas and Dill, 1996a].

Dans le cas de l'approximation quasi-chimique, on a vu que l'on posait que :

$$e_{ab} = \ln \frac{\rho_{ab}}{\rho_{ab}^{ref}}. \tag{2.24}$$

Thomas et Dill, quant à eux, suggèrent que les énergies e_{ab} doivent être telles que les fréquences des motifs structuraux observées dans la base de données (par exemple la fréquence ρ_{ab} des contacts entre a et b) doivent être égales aux fréquences avec lesquelles ces motifs structuraux seraient produits, en simulant des données sous la distribution de Boltzmann. La méthode itérative découle alors naturellement de ce principe : étant donné la valeur courante des énergies du potentiel, $e^{(n)}$, on calcule :

$$\Delta e_{ab}^{(n)} = -\ln \frac{\rho_{ab}(observé)}{\rho_{ab}(prédit)}, \tag{2.25}$$

où les valeurs prédites sont définies par :

$$\rho_{ab}(prédit) = \sum_s \sum_c \frac{n_{ab}^{c,p} \exp(-E(s,c))}{\sum_{c'} \exp(-E(s,c'))}, \tag{2.26}$$

où s représente une séquence protéique, c une structure, $E(s,c)$ l'énergie de la séquence s dans la structure c et $n_{ab}^{c,p}$ représente le nombre de contacts entre a et b prédits dans la structure c pour la protéine p. Et ensuite, on corrige le potentiel par :

$$e_{ab}^{(n)} = e_{ab}^{(n-1)} + \Delta e_{ab}^{(n-1)}, \tag{2.27}$$

où$\Delta e_{ab}^{(n)}$ représente l'erreur entre l'observé et le prédit pour les valeurs de $e^{(n)}$.

Les valeurs initiales de la méthode itérative peuvent être prises égales à celles obtenues en utilisant l'approximation quasi-chimique. A noter que, dans la mesure où la simulation prend en compte la connectivité de la chaîne, la méthode revient finalement à corriger de plus en plus finement le potentiel, par rapport à l'approximation quasi-chimique utilisée comme point de départ. De fait, les tests montrèrent que leur potentiel était extrêmement efficace sur le modèle de treillis testé, mais également dans le cas du *gapless threading* où ils retrouvaient la structure native dans 85 à 90 % des cas. Ils remarquèrent également que le critère de classement (ici, la moyenne de Boltzmann) pourrait également être représenté par le Z-score (cf. 2.2.6.3), car ils sont insensibles aux augmentations linéaires des énergies, et donc permettent de garder les énergies optimisées dans un domaine raisonnable. Des potentiels obtenus à l'aide de cette méthode furent plus tard comparés à l'approximation

quasi-chimique présentées par [Miyazawa and Jernigan, 1985,Betancourt and Thirumalai, 1999], et montrèrent une efficacité bien meilleure que les deux autres potentiels [Bastolla et al., 2001].

Il est intéressant de noter que la méthode de Thomas et Dill est très proche de ce que nous avons développé, à ceci près que nous avons travaillé dans un cadre de *protein design*. Dans notre cas, on applique le principe que les énergies du potentiel doivent être telles que les fréquences auxquelles les motifs structuraux (par exemple les fréquences ρ_{ab} des positions en contact qui présentent les acides aminés a et b) en simulant des ***séquences*** (et non des structures, comme dans le cas de Thomas et Dill) sous la distribution de probabilité induite par le modèle. De plus, on a montré dans ce cas que le principe était équivalent au principe du maximum de vraisemblance (cf. section 4.2). Il est très probable que cela soit également le cas pour la méthode de Thomas et Dill.

Cependant, la majorité des techniques d'optimisation de potentiels statistiques dans le cadre du *protein folding* ne sont pas basées sur de telles méthodes, mais cherchent à maximiser la différence d'énergie entre la structure native et les structures alternatives.

2.2.6.2 Programmation linéaire

Les algorithmes de programmation linéaire permettent d'optimiser une fonction sous un ensemble de contraintes, décrites par des inégalités. C'est un problème plus connu dans le domaine des mathématiques, dont la première approche pour résoudre des inégalités date de Fourier. La preuve que la résolution des inéquations peut se faire en un temps polynomial date cependant de 1979, rendant très séduisante l'idée de transformer un problème en un ensemble d'inéquations linéaires. Même si l'algorithme clef du simplex fut proposé dans les années 1950 [Dantzig et al., 1955], il existe d'autres algorithmes efficaces de résolution d'inéquations également utilisés dans ce contexte d'optimisation de potentiels dans le cadre du *protein folding* [Qiu and Elber, 2005].

Dans ce contexte, on cherche donc à optimiser la fonction d'énergie, ou plutôt ses paramètres, sous un ensemble de contraintes que l'on peut résumer simplement : la séquence native doit avoir la plus basse énergie quand elle est repliée dans sa structure native par rapport à toutes les autres structures possibles. Même si la puissance de calcul de ces dernières années a augmenté de manière exponentielle, il reste toujours impossible de définir de manière exacte toutes les structures possibles pour une séquence (excepté dans les modèles de treillis). L'on doit donc construire un jeu de structures alternatives pour chaque protéine appartenant au jeu d'apprentissage, puis résoudre, pour chaque protéine

dont la séquence est $\tilde{s}$, l'inégalité suivante créée par la contrainte thermodynamique :

$$E(\tilde{s}, \tilde{c}) < E(\tilde{s}, c') \qquad \forall c' \in \mathbb{C}, \tag{2.28}$$

où $\tilde{c}$ représente la structure native de la séquence $\tilde{s}$, et $\mathbb{C}$ l'ensemble des structures alternatives utilisées pour l'apprentissage. Pour une protéine donnée, il faut résoudre cette inéquation par rapport à toutes les structures alternatives du jeu de données, puis résoudre conjointement ces inéquations avec les autres groupes d'inéquations de chaque protéine du jeu d'apprentissage, menant à des centaines de milliers d'inégalités [Qiu and Elber, 2005].

Le premier potentiel obtenu par programmation linéaire [Maiorov and Crippen, 1992], était un potentiel de distance obtenu à partir d'un jeu de 37 protéines. Par la suite, la programmation linéaire [Vendruscolo and Domany, 1998b, Tobi and Elber, 2000, Micheletti et al., 2001, Loose et al., 2004] ou quadratique [Chhajer and Crippen, 2002] fut utilisée pour essayer de produire des potentiels statistiques, à l'aide de différents algorithmes comme la méthode du point intérieur [29].

Les systèmes d'inégalités décrits par l'équation (2.28) supposent que toutes les structures alternatives ont une énergie plus basse que la structure native. Cependant, il existe des structures natives de protéine qui ne sont pas les structures les plus stables [Sohl et al., 1998]. Plus trivialement, il existe de trop nombreuses violations de modèles pour qu'une telle propriété soit vérifiée par l'ensemble des protéines du jeu d'apprentissage, et il devient facile de trouver ne serait-ce qu'une seule structure dont l'énergie soit plus basse que la structure native. Ainsi, Vendruscolo et Domany montrèrent qu'un simple potentiel de contact ne pouvait pas être optimisé par programmation linéaire en respectant la contrainte exprimée à l'équation (2.28), même pour une unique protéine, la Crambin [Vendruscolo and Domany, 1998b]. Un peu plus tard, ils [Vendruscolo et al., 2000] présentèrent le résultat suivant : quel que soit le potentiel de contact choisi, il existe au moins une structure qui ne vérifie pas ces inégalités. D'un autre côté, Tobi et Eber parvenaient à la même conclusion, mais rajoutaient que, un cycle avant que le problème ne devienne non-résolvable (c'est à dire jusqu'à ce qu'on ne puisse plus résoudre les inégalités restantes), les potentiels obtenus avaient une précision qu'ils jugeaient satisfaisante (c'est à dire qu'il ordonnait correctement un sous-ensemble de structures alternatives) [Tobi and Elber, 2000]. Plus tard, la structure native fut remplacée dans l'algorithme par une structure approximée, représentant un mélange entre la structure native et les structures homologues [Qiu and Elber, 2005]. Le potentiel ainsi optimisé était relativement convergent (en partant de valeurs de potentiel aléatoires, les potentiel optimisés présentaient une efficacité semblable

29. Algorithme optimisé en 1984 par Narendra K. Karmarkar.

sur les jeux tests), et semblaient au moins aussi performants qu'un potentiel obtenu par estimation [30] ou optimisation du Z-score.

Un autre type de méthode se basant sur la programmation linéaire ou quadratique fut également construit en relaxant l'équation (2.28), c'est à dire en plaçant les structures natives et les structures homologues dans un bassin d'attraction d'énergie basse [Micheletti et al., 2001, Chhajer and Crippen, 2002]. Ainsi, cela permet notamment de lisser le paysage énergétique rugueux créé par les potentiels précédents [Tobi and Elber, 2000], tout en vérifiant que les structures natives ont une plus basse énergie que la plupart des autres structures, à condition que les structures alternatives ne soient pas homologues aux structures natives.

2.2.6.3 Optimisation du Z-score

Plutôt que d'essayer de trouver le potentiel qui minimise l'énergie de la structure native par rapport à toutes les structures alternatives, une autre technique consiste à chercher le potentiel qui maximise la différence d'énergie entre la structure native et la moyenne de l'énergie des structures alternatives. Cette différence d'énergie est facilement calculable par le Z-score :

$$Z = \frac{E_{nat} - \langle E_{leurre} \rangle}{\sigma_{leurre}(E)}, \tag{2.29}$$

où E_{nat} est l'énergie de la structure native, $\langle E_{leurre} \rangle$ l'énergie moyenne des structures alternatives, et $\sigma_{leurre}(E)$ la variance de l'énergie sur les structures leurres. Plus ce Z-score sera grand, plus la différence d'énergie entre la structure native et les structures leurres sera maximisée. Le terme $\sigma_{leurre}(E)$ permet d'empêcher que les potentiels n'amplifient artificiellement ce Z-score en augmentant simplement l'énergie des structures leurres les plus défavorables. Ce Z-score est beaucoup utilisé pour mesurer l'efficacité de potentiels, ou définir des structures leurres plus ou moins proches de la structure native.

En tant que critère d'optimisation, on lui préfère parfois une valeur dérivée. Par exemple, une première approche du Z-score en tant que critère d'optimisation fut définie par Goldstein et al, qui cherchaient à trouver le potentiel qui maximisait T_f/T_g, correspondant au ratio entre la température de repliement et celle de la transition vitreuse (méthode d'optimisation GLW). D'après leur définition de ces deux températures, cela revenait à maximiser le Z-score [Goldstein et al., 1992].

Cependant, la définition de la fonction de Z-scores qui peut prendre en compte toutes les protéines de la base de données reste un problème. Si l'on cherche simplement à trouver le potentiel qui maximise la somme des Z-scores, ce potentiel pourrait être optimal

30. Par exemple par l'utilisation de l'approximation quasi-chimique.

pour un certain nombre de protéine, et mauvais pour les autres. Ainsi, il faut trouver un moyen d'optimiser les potentiels selon le critère du Z-score pour toutes les protéines conjointement.

De leur côté, Shakhnovich et al cherchèrent à minimiser $\langle Z \rangle_{harm}$, décrite en tant que moyenne harmonique [Mirny and Shakhnovich, 1996, Chen and Shakhnovich, 2005] (méthode d'optimisation MS) :

$$\langle Z \rangle_{harm} = \frac{P}{\sum_{1 \leq p \leq P} 1/Z_p}, \tag{2.30}$$

où P est le nombre de protéines dans la base de données. Le test initié par [Thomas and Dill, 1996b] fut alors utilisé pour comparer les potentiels : on décrit dans un modèle de treillis, de toutes les structures possibles pour une séquence donnée, et on détermine la structure native d'après un potentiel p_1, posé comme le vrai potentiel, et enfin, on optimise un potentiel p_2 sur la base de données créée précédemment. A l'aide de ce test, ils montrèrent qu'une fois de plus le potentiel optimisé prédisait mieux les interactions favorables que les interactions défavorables (i.e. répulsifs), mais le potentiel p_2 présentait une bien meilleure corrélation avec le potentiel p_1 que des potentiels p_3 obtenus à l'aide de l'approximation quasi-chimique.

Plus intéressant du point de vue de cette thèse, on peut noter la méthode de Chiu et Goldstein, qui utilise la probabilité (exprimée en fonction du Z-score) de succès d'une séquence s une structure c, $P(\mathcal{S}(s))$ [Chiu and Goldstein, 1998b, Chiu and Goldstein, 2000] :

$$P(\mathcal{S}(s)) = \left(0,5 + 0,5 \text{ erf}\left[\frac{Z_s}{\sqrt{2}}\right]\right)^N, \tag{2.31}$$

où N représente le nombre de structures incorrectes pour cette séquence s (erf() étant la fonction d'erreur de Gauss). De ce fait, la probabilité $P(\mathcal{S}(s))$ représente la fonction de répartition de la loi normale, élevé à la puissance N. Afin de maximiser la probabilité de succès pour toutes les séquences de la base de données, il leur suffisait donc de maximiser $\langle P(\mathcal{S}(s)) \rangle$. Afin de mieux estimer les contacts répulsifs, ils définirent un poids, $\exp(-\alpha E_r)$, qu'ils associèrent à la contribution de chaque structure aléatoire r. Ce poids avait pour but de minimiser l'impact des structures ayant une énergie élevée. En comparant leur nouvelle méthode à différentes méthodes d'optimisation de la différence d'énergie, dont notamment GLW et MS, ils montrèrent que leur nouveau potentiel sous-estimait moins les contacts répulsifs et était plus efficace pour retrouver les repliements que les précédentes méthodes (taux de succès de 83 % par rapport à des taux de succès variant entre 53 et 68 %).

Une nouvelle fois, on peut voir cette dernière méthode comme une manière de maxi-

miser une vraisemblance, puisque l'on cherche à trouver le potentiel qui maximise la probabilité que l'on retrouve la structure native (les données), sachant le potentiel, la séquence native, et le jeu de structures, contenant la structure native et les structures alternatives (donc les paramètres du modèle).

2.2.6.4 *Overlap*

Une dernière méthode d'optimisation, très intéressante, a été développée par Bastolla et al. Dans leur groupe, ce terme de recouvrement (*overlap*) entre la matrice de contact native c et une matrice de contact d'une autre structure c', $q(c, c')$, apparaît d'abord comme une mesure permettant de comparer le recouvrement entre les matrices de contact prédite et native [Vendruscolo and Domany, 1998b] :

$$q(c, c') = \frac{N_{cc'}}{N_c}, \tag{2.32}$$

où $N_{cc'}$ est le nombre de contacts en commun dans les deux structures (prédite et native), et N_c est le nombre de contact dans la structure native. Le terme N_c permet de relativiser l'importance des contacts en communs par rapport à la taille de la protéine. La formule de l'*overlap* fut raffinée [Vendruscolo et al., 2000] :

$$q(c, c') = \frac{N_{cc'}}{\max(N_{c'}, N_c)}. \tag{2.33}$$

L'introduction de $N_{c'}$ permet de ne pas donner un poids trop fort à des matrices de contact c' représentant des structures très compactes et introduisant beaucoup plus de contacts que n'en ont en commun c et c'. L'*overlap* $q(c, c')$ prend des valeurs entre 0 et 1, et $q(c, c') = 1$ si et seulement si les deux matrices de contact c et c' sont égales. Ce terme d'*overlap*, utilisé à la base comme simple mesure de recouvrement, fut introduit dans la procédure d'optimisation plus tard, la même année [Bastolla et al., 2000]. A partir de l'équation (2.33), il est possible de retrouver l'*overlap* natif moyen :

$$Q(S, U) = \langle q(c, c') \rangle, \tag{2.34}$$

où U correspond aux potentiels que l'on cherche à optimiser, et la moyenne est calculée à l'aide de la loi de Boltzmann :

$$\langle q(c, c') \rangle = \frac{q(c, c') \exp(-E(c', s))}{\sum_{c''} \exp(-E(c'', s))}, \tag{2.35}$$

où $E(c, s)$ représente l'énergie de la séquence s repliée dans la structure c. Si la moyenne $\langle q(c, c') \rangle$ est proche de 1 alors la structure native est égale ou *est très proche* de son état

de plus basse énergie. Bastolla et al cherchèrent donc les potentiels maximisant cette moyenne [Bastolla et al., 2000,Bastolla et al., 2001]. Ils comparèrent ensuite leur méthode aux différentes méthodes présentées un peu plus haut, et notamment à la méthode utilisant la programmation linéaire. On peut noter que la méthode de l'*overlap* permet surtout de lisser le paysage énergétique (plutôt rugueux dans le cas de certains potentiels obtenus par programmation linéaire [Tobi and Elber, 2000]).

Les principales méthodes d'optimisation de potentiels statistiques, dans le contexte du *protein folding* se rapportent aux quatre méthodes que je viens de citer. Ainsi l'on cherche en général les potentiels qui maximisent le contraste entre l'énergie des structures natives et chacune des autres structures, ou bien ceux qui maximisent l'énergie de la structure native par rapport à une une valeur moyenne d'énergie sur les autres structures.

2.3 Le problème du *protein design*

2.3.1 Introduction

L'approche historique des relations entre les séquences protéiques et leurs structures est celle du *protein folding*, c'est à dire que l'on cherche à retrouver la structure (repliement ou conformation) native d'une séquence donnée. Comme nous venons de le voir dans la section précédente, de très nombreuses méthodes ont été développées afin de trouver des potentiels statistiques (mais également des fonctions d'énergie semi-empiriques) adaptés à ce problème. D'un autre côté, il a souvent été observé que des séquences très différentes peuvent avoir le même repliement. Ainsi, le *protein design*, aussi appelé *inverse protein folding* considère le problème dans l'autre sens : si l'on connait un repliement, est-il possible de retrouver la séquence ou l'ensemble de séquences qui lui correspondent ?

Une protéine dispose d'un outil pour accomplir sa fonction : sa structure. Elle lui permet, par exemple, d'effectuer une réaction chimique (pour une enzyme par exemple) ou de reconnaître un ligand afin de générer une réaction appropriée de la cellule (récepteurs). Les protéines jouent un rôle majeur dans les organismes, sous la forme de récepteurs, d'enzymes, d'hormones, régulateurs, anticorps... Il est possible que l'on souhaite reconstruire une protéine d'intérêt ayant un repliement particulier, par exemple dans le domaine médical. Cependant, le nombre de séquences possibles pour une structure donnée est bien trop grand pour qu'il soit possible de générer toutes les séquences admissibles pour chaque repliement d'intérêt, et les tester expérimentalement. Aussi, l'intérêt du *protein design* pourrait être ici de réduire la quantité des séquences possibles, voire même dans certains cas de définir quelques séquences plus probables, à tester en priorité *in vitro*.

Les structures des protéines évoluent lentement au cours du temps [Chothia and Lesk, 1986], au contraire des séquences (nucléotidiques et protéiques), qui sont soumises à de fortes variations. De plus, les structures des protéines exercent une forte contrainte sur les séquences, afin que la protéine puisse continuer à accomplir sa fonction. Les séquences évoluent donc sous la contrainte d'une structure (presque) constante et on peut dire que l'évolution suit un processus de *protein design*. Dans le cadre qui nous intéresse, c'est à dire construire un modèle d'évolution moléculaire soumis à des contraintes structurales, l'utilisation d'une fonction d'énergie spécialement construite pour retrouver le repliement d'une séquence ne semble pas le plus adapté. Au contraire, l'approche de *protein design*, semble bien plus attractive, d'autant plus que les formulations mathématiques des deux problèmes induisent des approximations qui sont différentes selon l'approche considérée.

Il existe une littérature très importante sur le problème du *protein design*, en particulier dans une perspective où des champs de force semi-empiriques, définis à une résolution atomique, sont utilisés. Plusieurs méthodes ont été développées pour résoudre le problème du positionnement des chaînes latérales : par *dead-end elimination* [Desmet et al., 1992, Desjarlais and Clarke, 1998] ou par champ moyen [Koehl and Delarue, 1994, Koehl and Delarue, 1996]. Ces méthodes sont probablement bien plus précises que tout ce que nous avons pu faire de notre côté, mais a contrario, elles sont pour nous prohibitives en temps de calcul. Par conséquent, dans notre cas, nous nous appuierons plutôt sur la littérature basée sur des potentiels statistiques *coarse-grained*. De plus, dans ce qui suit, nous nous focaliserons avant tout sur la question plus théorique de comment *définir* correctement le critère d'optimisation, en particulier, eût égard au facteur de normalisation Z_s.

2.3.2 *Protein design* versus *protein folding*

Les potentiels statistiques sont construits de manière à vérifier l'équation de Boltzmann :

$$p(c|s) = \frac{\exp\left(-E(s,c)\right)}{\sum_{c' \in \mathbb{C}} \exp\left(-E(s,c')\right)} = \frac{\exp\left(-E(s,c)\right)}{Z_s}, \tag{2.36}$$

où $p(c|s)$ est la probabilité que la séquence s se replie dans la structure c, $E(s,c)$ est l'énergie de s associée à c d'après le potentiel statistique et $\mathbb{C}$ représente l'ensemble des structures.

Dans une approche de *protein folding*, pour une séquence s donnée, on cherche la structure $\hat{c}$ telle que :

$$p(\hat{c}|s) > p(c|s) \qquad \forall c \in \mathbb{C}, \tag{2.37}$$

que l'on peut exprimer de manière plus simple :

$$\begin{aligned} \exp\left(-E(s,\hat{c})\right)/Z_s &> \exp\left(-E(s,c)\right)/Z_s \\ \Leftrightarrow \exp\left(-E(s,\hat{c})\right) &> \exp\left(-E(s,c)\right) \\ \Leftrightarrow E(s,\hat{c}) &< E(s,c). \end{aligned} \tag{2.38}$$

Ainsi, dans une approche de *protein folding*, on va simplement rechercher la structure $\hat{c}$ qui minimise l'énergie.

Dans le cas du *protein design*, le but est de retrouver la séquence ou l'ensemble de séquences correspondant à une structure c donnée. Cependant, il ne suffit pas de trouver une séquence qui minimise l'énergie conjointe de la structure et de la séquence (comme on vient de le voir), mais il faut également que la séquence que l'on a retrouvée puisse se replier dans cette structure c. Ainsi, alors que le *protein folding* ne demande qu'à rechercher une structure adéquate parmi l'ensemble des structures, le *protein design* demande à ce que soit effectuée une recherche dans l'espace conjoint des structures et des séquences. De manière plus formelle, on peut définir que la condition mathématique du *protein design* est de trouver une séquence $\hat{s}$ telle que :

$$p(c|\hat{s}) > p(c|s) \qquad \forall s \in \mathbb{S}, \tag{2.39}$$

où $\mathbb{S}$ représente l'ensemble des séquences. Ceci se traduit de la manière suivante :

$$\frac{\exp\left(-E(\hat{s},c)\right)}{Z_{\hat{s}}} > \frac{\exp\left(-E(s,c)\right)}{Z_s} \tag{2.40}$$

$$\Leftrightarrow E(\hat{s},c) + \ln Z_{\hat{s}} < E(s,c) + \ln Z_s \tag{2.41}$$

$$\Leftrightarrow E(\hat{s},c) - F(\hat{s}) < E(s,c) - F(s) \tag{2.42}$$

où $F(s) = \ln Z_s$ est appelé énergie libre de la séquence s. Ainsi, contrairement au *protein folding*, le facteur de normalisation Z_s, qui dépend de la séquence, ne peut pas se simplifier, puisque s et $\hat{s}$ sont différentes. Or, ce facteur Z_s suppose une somme sur l'espace des structures. Il faut donc trouver un moyen d'approximer ce facteur de normalisation, pour une séquence donnée. Pour cela, deux approximations sont possibles : le *random energy model* et l'utilisation de structures alternatives. Comme nous avons vu à la section 2.1.2.2 plusieurs méthodes de construction de jeux de structures alternatives, je ne définirai ici que l'approximation induite par le *random energy model* avant de détailler quelques méthodes d'optimisation de potentiels dans le contexte du *protein design*.

2.3.3 Le *random energy model*

Une première hypothèse simpliste serait de considérer que, quelle que soit la séquence s, l'énergie libre de la séquence serait constante. Il suffirait alors de minimiser l'énergie du couple (c, s) en effectuant une recherche parmi l'ensemble des séquences. Cependant, même en travaillant avec des valeurs de potentiel constantes, cette première hypothèse mènerait à des séquences absurdes, formant par exemple des polycystéines (si l'on considère que l'interaction cystéine-cystéine est la plus favorable) [Shakhnovich and Gutin, 1993]. Une manière de contrer cet inconvénient est alors de travailler à composition constante, mais cela ne permet pas d'explorer l'ensemble des séquences possibles. Afin de pouvoir prendre en compte l'ensemble des séquences d'une même taille n, il suffit alors de simplement déterminer un facteur de normalisation ne dépendant que de la composition des séquences :

$$F(s) = \sum_{1 \leq i \leq n} \lambda_{s_i}, \tag{2.43}$$

où s_i représente l'acide aminé au site i. Cette approximation est appelée le *random energy model* (REM) [Shakhnovich and Gutin, 1993,Sun et al., 1995,Finkelstein et al., 1995a,Seno et al., 1998].

Posons s et s' deux séquences de taille n. Si on choisit d'ignorer la connectivité de la chaîne principale, alors le facteur de normalisation ne dépend que de la différence de composition des deux séquences. Si cette approximation est simple et facilement intégrable dans une procédure d'optimisation de potentiels, elle est souvent critiquée car, au même titre que l'approximation quasi-chimique de Miyazawa et Jernigan, il s'agit une approximation grossière, qui ne semble a priori pas représenter correctement l'ensemble des structures.

2.3.4 Optimisation versus échantillonnage

A noter que, comme le montre la diversité des séquences naturelles pour une conformation donnée, il n'est pas forcément adapté d'envisager le problème du *protein design* comme une optimisation (i.e. trouver $\hat{s}$ qui minimise $E(s, c) - F(s)$). On peut éventuellement obtenir une collection de séquences par seuillage, c'est à dire trouver les séquences $\hat{s}$ telle que :

$$E(\hat{s}) - F(\hat{s}) < \varphi. \tag{2.44}$$

Dans la suite, la formulation des modèles d'évolution soumis à des contraintes structurales permet de poser le problème en terme d'échantillonnage. Ainsi, on va échantillonner s

d'après la distribution de probabilité $p(s)$:

$$s \sim p(s) = \frac{e^{-(E(s,c)-F(s))}}{Y_c}, \tag{2.45}$$

où

$$Y_c = \sum_{s \in \mathbb{S}} e^{-(E(s,c)-F(s))}. \tag{2.46}$$

Cette méthode d'échantillonnage, particulièrement utilisée dans le modèle *SC* et dans le chapitre 4, est décrite dans le chapitre 3.

2.3.5 Optimisation de potentiels dans un contexte de *protein design*

On a vu précédemment que l'approche du *protein design* et celle du *protein folding* étaient différentes. Les potentiels obtenus dans le contexte du *protein folding* donnent de bons résultats pour retrouver les repliements des protéines. Par exemple, la méthode présentée dans [Chiu and Goldstein, 2000] présentait un taux de succès de 83 % dans un modèle de treillis. A l'inverse, les potentiels utilisés dans le cadre du *protein design* semblent beaucoup moins performants. Et donc, ce domaine propose des challenges intéressants. En outre, les protéines semblent évoluer sous une contrainte de *protein design*, et non de *protein folding*, puisque les structures évoluent lentement, et qu'elles exercent une contrainte forte sur les séquences. Ainsi, se tourner vers l'optimisation de potentiels statistiques dans un contexte de *protein design* semble particulièrement attirant.

Les méthodes permettant d'optimiser des potentiels statistiques explicitement pour le problème du *protein design* ont été beaucoup moins développées que celles pour le *protein folding*. On utilise encore l'approximation quasi-chimique pour obtenir des potentiels statistiques [Dehouck et al., 2006] afin de répondre aux problèmes du *protein folding* et du *protein design*. Comme nous recherchions un potentiel statistique explicitement optimisé dans le cadre du *protein design*, afin de l'intégrer dans un modèle d'évolution soumis à des contraintes structurales, nous avons décidé de créer notre propre méthode d'optimisation, qui sera décrite dans la partie suivante. De plus, nous cherchions également à définir un cadre statistique complet nous permettant de tester différentes méthodes d'optimisation.

Les méthodes déjà existantes consistent à optimiser une fonction de forme variable, en essayant de prendre en compte le mieux possible le facteur de normalisation, le problème résidant dans la manière d'explorer les deux espaces de recherche.

Parmi les méthodes utilisant l'approximation du *random energy model*, on peut citer l'optimisation proposée par Seno et al, qui consiste à optimiser une fonction Δ telle

que [Seno et al., 1998] :

$$\Delta = \bar{\epsilon}\left[\sum_k \left(\frac{E(s_k,c_k)-F(s_k)}{L_k}\right)^2 + \left(\frac{E(s_k,c_k)-F(s_k)}{L_k}\right)^4 \Theta\left(F(s_k)-E(s_k,c_k)\right)\right], \tag{2.47}$$

où L_k est la longueur de la k-ième séquence, c_k la conformation native de la protéine k, et

$$\Theta(F(s_k)-E(s_k,c_k)) = \begin{cases} 0 & \text{si } F(s_k) < E(s_k,c_k) \\ 1 & \text{sinon} \end{cases}. \tag{2.48}$$

Cette fonction permet notamment de pénaliser les potentiels qui placerait l'énergie de la protéine à des valeurs physiquement impossibles. Cette fonction est minimisée à l'aide d'une procédure de recuit simulé, en forçant les potentiels à respecter un ordonnancement des forces d'attraction, déduites de la base de données (c'est à dire afin que les interactions les plus présentes dans la base de données soient les plus favorables).

Une autre méthode d'optimisation fut décrite par Deutsch et Kurosky. À l'aide d'une fonction à minimiser ΔF telle que [Deutsch and Kurowski, 1996, Deutsch and Kurowski, 1997] :

$$\Delta F = \sum_{1\leq k\leq N} E(c_k^*, s_k, P) - F(s_k, P), \tag{2.49}$$

où E représente le potentiel statistique, P le jeu de paramètres du potentiel, et $F(s_k, P)$ correspond à l'énergie libre la séquence s_k. A partir d'un premier ensemble de valeurs du potentiel, pour chaque séquence, des structures peuvent être échantillonnées à l'aide d'une procédure de recuit simulé afin de calculer l'énergie libre de la séquence. Cette méthode fut appliquée dans un modèle de treillis, où les structures sont très faciles à générer, et dans un modèle réel de protéine, pour un ensemble de 12 structures avec une taille de 8 acides aminés [Deutsch and Kurowski, 1997]. Cette technique est très attrayante, mais elle demande malheureusement de générer un ensemble de structures en fonction d'un potentiel, ce qui est difficile actuellement pour des bases de données réelles.

De leur côté, Chiu et Golstein utilisèrent la même forme de probabilité qu'ils avaient utilisée pour trouver des potentiels pour le problème du *protein folding* (cf. 2.2.6.3), mais en définissant non plus le score d'une séquence par rapport aux nouveaux potentiels, mais le score d'une structure [Chiu and Goldstein, 1998a] :

$$P(\mathcal{S}(c)) = \left(0,5 + 0,5 \text{ erf}\left[\frac{Z_c}{\sqrt{2}}\right]\right)^N. \tag{2.50}$$

L'ensemble des structures étaient entièrement décrit à l'aide d'un modèle de treillis de taille 3^3 pour des séquences constituées de 27 acides aminés. Chiu et Goldstein montrèrent également que, à l'aide d'une telle approche (et en se basant sur le test présenté dans [Thomas and Dill, 1996b]), le potentiel optimisé générait des séquences plus proches des séquences natives (pour un jeu de données indépendant du jeu d'apprentissage) que le véritable potentiel lui-même.

Dans ce chapitre, j'ai résumé plusieurs manières d'essayer de prendre en compte la dépendance entre la séquence et la structure d'une protéine au travers d'une fonction. Bien que la manière la plus attrayante de construire une telle fonction, dans le cadre qui nous intéresse, serait d'utiliser des champs de force semi-empiriques exprimant les relations entre les atomes des protéines, nous nous sommes plutôt tournés vers l'optimisation de potentiels statistiques. En effet, ceux-ci présentent, dans notre contexte, de nombreux avantages par rapport aux champs de force semi-empiriques, et notamment, pour la forme de potentiel simplifié choisie, de pouvoir être calculés très rapidement. Au sein des potentiels statistiques, on peut distinguer deux types d'approches, l'une visant à retrouver la structure d'une séquence donnée (*protein folding*) et l'autre cherchant un ensemble de séquences correspondant à une structure fixée (*protein design*). A chaque approche sont associées plusieurs méthodes d'optimisation, mais les méthodes d'optimisation de potentiels de *protein design* sont en général moins développées que celles pour le *protein folding*. Comme le modèle d'évolution moléculaire soumis à des contraintes structurales semble s'insérer dans un cadre de *protein design*, nous avons donc décidé de créer notre propre cadre de travail, et d'ainsi d'obtenir des potentiels qui seraient consistants avec le modèle d'évolution moléculaire. Ceci est donc le sujet de cette thèse, et le travail effectué dans ce but sera décrit dans la partie II. Cependant, avant d'entrer dans le vif du sujet, je décrirais les méthodes numériques et statistiques utilisées dans cette thèse.

Chapitre 3

Méthodes numériques et statistiques

3.1 Introduction et notations

Cette thèse fait appel à plusieurs méthodes numériques et statistiques, qui sont développées dans les articles présentés dans le cadre de ce mémoire ou dans d'autres articles connexes. Cependant, il est intéressant de regrouper ces différentes méthodes au sein d'un chapitre introductif, afin de poser le cadre statistique ayant servi de base au travail effectué. Ces méthodes seront expliquées à l'aide des variables utilisées dans le modèle d'évolution et dans l'optimisation des paramètres du potentiel statistique.

L'on notera s une séquence protéique de taille n, et c sa structure native associée. Θ correspond au jeu de paramètres associés au modèle d'évolution M (et Ξ représente l'espace des paramètres, $\Theta \in \Xi$), et θ fait référence au jeu de paramètres du potentiel statistique E (énergies de contact, d'accessibilité au solvant). D correspond aux données, représentant soit un alignement de séquences nucléotidiques dans le contexte du modèle d'évolution, soit des protéines (couples séquence-structure) issues de la PDB dans le contexte de l'optimisation des paramètres.

La première des méthodes numériques présentées ici est utilisée intensivement dans le cadre du modèle d'évolution. Il s'agit d'échantillonner Θ à l'aide de l'algorithme de Metropolis-Hasting (MH). La deuxième méthode consiste à échantillonner des séquences s suivant une distribution de probabilité, $p(s|c,\theta)$, à l'aide de l'algorithme d'échantillonnage de Gibbs (GS), une méthode qui sera notamment utilisée dans le chapitre 4. La troisième méthode correspond à la descente de gradient afin de trouver un jeu de paramètres θ optimal, sous certaines conditions. Ces deux méthodes (échantillonnage de Gibbs et descente de gradient) sont également décrites dans les articles suivants, mais on souhaite ici faire une récapitulation technique des méthodes utilisées dans cette thèse afin

de les relier entre elles. La dernière méthode utilisée ici est le calcul du facteur de Bayes pour comparer deux modèles d'évolution, décrit plus particulièrement dans les articles de Rodrigue et al (cf. [Rodrigue et al., 2006] pour une description détaillée de la méthode).

3.2 Algorithme de Metropolis-Hasting

Dans le modèle d'évolution, qui se situe dans un contexte Bayésien, on est souvent amené à considérer la probabilité postérieure d'un jeu de paramètres $\Theta \in \Xi$, sachant un modèle M, $p(\Theta|D, M)$, définie à l'aide du théorème de Bayes :

$$p(\Theta|D, M) = \frac{p(D|\Theta, M)p(\Theta|M)}{p(D|M)}, \tag{3.1}$$

où $p(\Theta|D, M)$ représente la probabilité a priori du jeu de paramètre Θ pour ce modèle M, et $p(D|\Theta, M)$ est la vraisemblance. La vraisemblance marginale, $p(D|M)$ est exprimée par :

$$p(D|M) = \int_{\Xi} p(D|\Theta, M)p(\Theta|M)d\Theta. \tag{3.2}$$

La plupart des estimateurs Bayésiens reviennent à calculer des espérances sur la probabilité à posteriori. De telles espérances ne sont en général pas analytiquement calculables. Cependant, en statistique, il est possible d'approximer de telles espérances à partir d'un échantillon de valeurs de Θ tirées de $p(\Theta|D, M)$, à condition que cet échantillon soit assez grand. Par exemple, en supposant que la mesure de K états soit suffisante, on peut estimer l'espérance de Θ, $\langle\Theta\rangle$,de la manière suivante :

$$\langle\Theta\rangle = \int_{\Xi} \Theta p(\Theta|D, M)d\Theta \tag{3.3}$$

$$\simeq \frac{1}{K} \sum_{1\leq k\leq K} \Theta^{(k)}, \tag{3.4}$$

où l'on a échantillonné chaque $\Theta^{(k)}$ de manière à ce que $\Theta^{(k)} \sim p(\Theta|D, M)$. A partir de l'équation (3.4), il est alors possible de calculer numériquement les valeurs d'intérêts dépendant de Θ, mais pas l'équation (3.2), comme nous le verrons à la section 3.5.

Afin d'échantillonner les K valeurs $\Theta^{(k)} \sim p(\Theta|D, M)$, on fait ici appel aux chaînes de Markov Monte Carlo (MCMC) et à l'algorithme MH [Metropolis et al., 1953, Hastings, 1970]. Il s'agit de générer une chaîne de Markov à valeurs dans Ξ , $(\Theta^{(k)})_{k\in\mathbb{N}}$, dont la distribution d'équilibre est justement $p(\Theta|D, M)$. Pour ce faire, à supposer que la valeur courante soit $\Theta^{(k)} = \Theta$, alors on propose un état candidat, nommé Θ^*, d'après un noyau

stochastique, tel que $q(\Theta, \Theta^*)$ est la densité de probabilité de proposer Θ^* à partir de Θ. Celle-ci doit être choisie de manière à être facilement implémentable et calculable analytiquement [Neal, 1993]. On accepte ensuite le nouveau jeu de paramètres Θ^*, c'est à dire qu'alors $\Theta^{(k+1)} = \Theta^*$, avec une probabilité $A(\Theta, \Theta^*)$:

$$A(\Theta, \Theta^*) = min \left\{ 1, \underbrace{\frac{p(\Theta^*|D, M)}{p(\Theta|D, M)}}_{\text{ratio de Metropolis}} \cdot \underbrace{\frac{q(\Theta^*, \Theta)}{q(\Theta, \Theta^*)}}_{\text{ratio de Hasting}} \right\}, \tag{3.5}$$

sinon on rejette le jeu de paramètres θ^* et $\Theta^{(k+1)} = \Theta$. Le ratio de Hasting permet de corriger la probabilité d'acceptation si les densités de proposition sont asymétriques. De cette manière, on construit une suite $\Theta^{(k)}$, qui est une chaîne de Markov dont la distribution stationnaire est $p(\Theta|D, M)$. De plus, la convergence vers la distribution stationnaire est géométrique. En pratique, il faut laisser "tourner la chaîne" (itérer l'algorithme) pendant un nombre de cycles parfois élevé, et en tout état de cause très difficile à estimer a priori, avant de voir la chaîne parvenir à son état d'équilibre. Cette partie préparatoire, qui n'est pas prise en compte dans l'échantillonnage, est appelée *burn-in*. A partir du moment où la chaîne a atteint la stationnarité, on échantillonne les $\Theta^{(k)}$ utilisés pour l'estimation dans l'équation (3.4) à intervalles réguliers.

3.3 Algorithme d'échantillonnage de Gibbs

Une autre méthode d'échantillonnage utilisée par cette thèse est l'algorithme d'échantillonnage de Gibbs. Plus simple que l'algorithme MH, celui-ci nous sert ici à échantillonner des séquences s d'après la distribution de probabilité $p(s|c, \theta)$. Pour obtenir un échantillon $s^{(k)}$, l'on cherche à construire une chaîne de Markov mais cette fois-ci à valeurs dans $\mathbb{S}$. En l'occurrence, on cherche à calculer $\langle f(s, c|\theta) \rangle$, l'espérance d'une fonction qui dépend de s et de c, que l'on peut définir de la même manière qu'à l'équation (3.4) :

$$\langle f(s, c|\theta) \rangle = \int_{\mathbb{S}} f(s, c|\theta) p(s|c, \theta) ds \simeq \frac{1}{K} \sum_{1 \leq k \leq K} f(s^{(k)}, c|\theta), \tag{3.6}$$

où $s^{(k)} \sim p(s|c, \theta)$.

L'algorithme de Gibbs consiste à prendre les sites de la séquence un à un ($i = 1..N$), et, pour chacun, réévaluer l'acide aminé en cette position, conditionnellement au reste de la séquence en toutes les autres positions. Supposons que l'on réévalue la position i. Alors, dans un premier temps, on calcule, pour chaque $a = 1..20$ (i.e. pour chaque acide aminé

possible en i), la probabilité conditionnelle d'avoir l'acide aminé a en la position i :

$$g_i(a) = p(s_i = a|s_{\setminus i}, c). \tag{3.7}$$

Ici, $s_{\setminus i}$ désigne tout le reste de la séquence (hormis la position i). Cette probabilité somme à un sur tous les acides aminés :

$$\sum_{1 \leq a \leq 20} g_i(a) = 1. \tag{3.8}$$

Ensuite, on tire un acide aminé d'après cette distribution de probabilité (par exemple a^*) et l'on pose $s_i = a^*$. Enfin, on itère sur toutes les positions i. En prenant successivement tous les sites de la séquence, on aura alors échantillonné une nouvelle séquence $s^{(n)}$ de sorte que la chaîne de Markov $(s^{(n)})$, pour $n > 0$, a pour distribution stationnaire $\sim p(s|c, \theta)$, et on considèrera avoir effectué une itération. On peut noter que l'échantillonnage de Gibbs permet d'échantillonner de nouvelles valeurs pour chaque variable aléatoire (ici, les acides aminés à chaque site) étant donné des valeurs fixées pour les autres variables aléatoires (tous les autres sites différents du site échantillonné sont considérés comme fixés).

De même que pour l'algorithme MH, il faut effectuer un certain nombre d'itérations avant que la chaîne n'atteigne sa stationnarité. De plus, il faut effectuer un certain nombre d'itérations entre chaque séquence choisie. Cette méthode est utilisée dans le chapitre 4, afin de calculer le gradient sur les paramètres θ.

Avant de détailler la méthode du gradient, intensivement utilisée au cours de cette thèse, il est intéressant de remarquer que les méthodes d'échantillonnage par MCMC ne favorisent pas forcément l'état le plus probable, mais autorisent une certaine "marge d'erreur", dans le sens où il est toujours possible de choisir des états ayant une probabilité non maximale. Des aspects combinatoires non triviaux interviennent également ici.

3.4 Méthode de descente de gradient

La méthode de la descente de gradient est la méthode la plus ancienne d'optimisation sans contraintes. Soit $f(\mathbf{x})$ une fonction continue qui dépend du vecteur $\mathbf{x}$, de dimension P, tel que $\mathbf{x} = \{x_0, x_1, ..., x_P\}$, $x_p \in \mathbb{R}$. $f(\mathbf{x})$ prend des valeurs réelles, avec un optimum global en $\mathbf{X}^*$, que l'on cherche à déterminer. L'optimum est ici un minimum, et donc, formellement, on cherche :

$$\mathbf{X}^* = argmin(f). \tag{3.9}$$

Prenons un vecteur au hasard, $\mathbf{x}^0$, que l'on sait ne pas être l'optimum. Soit la dérivée

partielle de f par rapport à la variable x_p,

$$\frac{\partial f}{\partial x_p} \neq 0, \tag{3.10}$$

qui est par définition la pente de f par rapport à x_p. Donc la direction du minimum dans la dimension p est indiquée par $-\frac{\partial f}{\partial x_p}$. Afin d'obtenir un point meilleur que $\mathbf{x}^0$ dans la dimension p, il suffit de faire un déplacement (pas) dans la direction de $-\frac{\partial f}{\partial x_p}$. Cependant, la taille du pas est critique : si elle est trop élevée, on risque de passer au delà de l'optimum, et si elle est trop faible, la descente de gradient sera très lente. Formellement, la descente de gradient se caractérise par :

$$\mathbf{x}^t = \mathbf{x}^{t-1} - \delta \frac{df}{d\mathbf{x}}, \tag{3.11}$$

où δ représente le pas de la descente de gradient. On appelle Δ^t le pas effectué entre $t-1$ et t :

$$\Delta^t = -\delta \frac{df}{d\mathbf{x}}. \tag{3.12}$$

L'équation (3.11) constitue la plus simple des méthodes de descente de gradient, et le gradient s'annule au voisinage de l'optimum.

Le problème d'une telle descente de gradient est l'existence, pour certaines fonctions f, d'optima locaux. En ces points, de même que dans l'optimum global, le gradient disparaît :

$$\frac{df}{d\mathbf{x}} = 0, \tag{3.13}$$

et les gradients alentours mènent à cet optimum local (cf. figure 3.1).

Une méthode permet parfois de sortir de certains optimums locaux, même s'il ne s'agit pas de sa justification première : la méthode du gradient inertiel, inspiré de l'inertie physique. Cette méthode fut développée afin d'accélérer la descente de gradient. Il s'agit d'un gradient "à mémoire", c'est à dire que, à chaque pas, on ajoute aux nouvelles valeurs de $\mathbf{x}$ une composante qui dépend des pas effectués auparavant :

$$\begin{aligned} \Delta^t &= -\delta \frac{df}{d\mathbf{x}} \\ \mathbf{x}^t &= \mathbf{x}^{t-1} + \Delta^t + \delta_{iner} \Delta^{t-1}, \end{aligned} \tag{3.14}$$

où $0 \neq \delta_{iner} < 1$. Quand $\delta_{iner} = 0$, on retrouve une descente de gradient simple. Une autre méthode de descente de gradient utilisée dans cette thèse est décrite dans le chapitre 5.

Malgré tout, ces méthodes ne garantissent pas que les valeurs des variables optimisées, $\mathbf{x}^\bullet$, soit le jeu de valeurs correspondant à l'optimum global, $\mathbf{X}^*$. Cependant, si l'on réalise

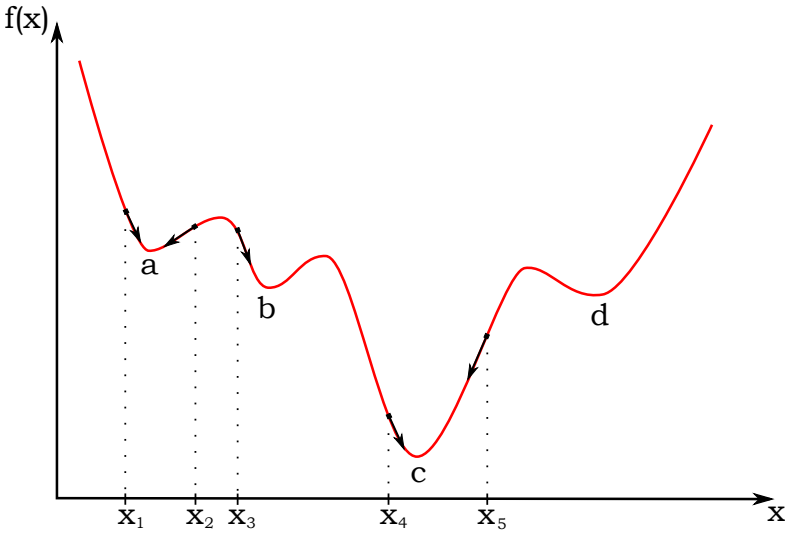


FIGURE 3.1 – Une fonction $f(x)$ présentant un minimum global (c) et des minimums locaux (a, b, d). Les flèches représentent les directions du gradient pour chaque point $x_k, k \in \{1..5\}$.

un grand nombre d'optimisations à partir de valeurs aléatoires pour toutes les variables de $\mathbf{x}$, et que les jeux de valeurs sont identiques (tout comme la valeur de l'optimum trouvé), l'on peut raisonnablement supposer qu'il s'agit des valeurs correspondant à l'optimum global [31].

Dans le cadre de l'optimisation des paramètres, $f(x) = -\ln P(D|\theta, M)$, et les variables $\mathbf{x}$ sont les composantes du potentiel statistique, θ. L'on réalisera aussi un grand nombre d'optimisations à partir de valeurs aléatoires du potentiel afin de vérifier que l'optimum trouvé n'est pas un minimum local.

3.5 Facteur de Bayes

Le facteur de Bayes nous permet de comparer deux modèles (ici, deux modèles d'évolution). Le facteur de Bayes entre deux modèles M_0 et M_1 est défini comme étant le rapport des vraisemblances marginales :

$$B_{01} = \frac{p(D|M_1)}{p(D|M_0)}, \tag{3.15}$$

31. Ou bien que l'optimum local trouvé a un bassin d'attraction bien plus large et profond que l'optimum global.

où $p(D|M_i)$ est l'intégrale définie à l'équation (3.2) [Jeffreys, 1935]. Le facteur de Bayes pénalise intrinsèquement les modèles ayant le plus de paramètres par rapport aux modèles simples.

Le facteur de Bayes est notoirement difficile à évaluer numériquement. Une méthode assez complexe, mais fiable, consiste à effectuer une intégration thermodynamique : cette méthode consiste à tracer un chemin entre deux modèles, par de subtiles variations entre le/les paramètre(s) différant entre les deux modèles. Par exemple, supposons que les deux modèles ne diffèrent que par un paramètre, β, tel que $\beta = 0$ pour M_0 et $\beta = x$ pour M_1.

Par définition,

$$\ln B_{01} = \ln p(D|M_1) - \ln p(D|M_0) \tag{3.16}$$

$$= \int_0^x \frac{\partial \ln p(D|M_\beta)}{\partial \beta} d\beta \tag{3.17}$$

or

$$\frac{\partial \ln p(D|M_\beta)}{\partial \beta} = E_\Theta \left[\frac{\partial \ln p(D|\Theta, M_\beta)}{\partial \beta} \right] \tag{3.18}$$

où l'espérance est prise sur la distribution a posteriori sous le modèle $M_\beta : \Theta \sim p(\Theta, D, M_\beta)$. Et donc, d'après [Rodrigue et al., 2006, Rodrigue, 2007], le facteur de Bayes peut être estimé de la manière suivante :

$$\ln B_{01} \simeq \left(\sum_{1 \leq k \leq K} \frac{\partial \ln p(D|M_{\beta_k})}{\partial \beta_k} \right) \cdot \frac{\delta\beta}{I} \tag{3.19}$$

où (β_k) est une suite régulière sur $(0, x)$ et $\Theta(k)$ est une chaîne de Markov Monte Carlo quasi-statique (obtenue en appliquant l'algorithme MH, mais en changeant β à chaque cycle). $\delta\beta$ est le pas qui sépare deux variations de β ($\delta\beta = \beta_{h+1} - \beta_h$), I la taille de l'intervalle (soit x) et X le nombre de pas séparant $\beta = 0$ de $\beta = x$. En variant β de 0 vers x et de x vers 0, on obtient deux estimateurs indépendants du facteurs de Bayes. Il est à noter que plus la valeur du pas $\delta\beta$ sera petite, plus le calcul sera précis.

N'importe quel chemin (continu et différentiable) peut être en principe choisi entre les deux modèles. Dans notre cas, et pour anticiper sur les chapitres à venir, un chemin naturel consiste à faire varier la stringence appliquée au potentiel statistique. Plus précisément, le modèle d'évolution moléculaire SC entre deux codons c et c' est défini par la matrice de substitution suivante :

$$R_{cc'} = \begin{cases} Q^{mut}_{bb'} \cdot e^{\frac{\beta}{2}(H(s|\gamma) - H(s'|\gamma))} & \text{si } a' \neq a \\ Q^{mut}_{bb'} & \text{si } a' = a \\ 0 & \text{si } c \text{ et } c' \text{ ne sont pas plus proches voisins} \end{cases}, \tag{3.20}$$

avec b représentant le nucléotide de la séquence σ muté en b' dans la séquence σ', a (resp. a') est l'acide aminé codé par c (resp. c'), et s (resp. s') est la séquence codée par σ (resp. σ'), et H est la fonction de sélection qui dépend du potentiel statistique E. Dans ce cas, le paramètre β représente un facteur de pondération du potentiel statistique, de telle sorte que lorsque $\beta = 0$, le modèle se réduit au modèle de référence M_0 (purement mutationel, sans sélection). En effectuant une intégration thermodynamique le long du chemin défini par la variation continue de β sur un intervalle issu de zéro, on peut "dérouler", pour ainsi dire, une courbe de fit du modèle, en fonction de β. Dans la figure 3.2, j'ai représenté l'évolution du facteur de Bayes pour le jeu de données GLOBIN pour les potentiels optimisés dans le chapitre 5. On voit sur la figure 3.2 que, pour le potentiel considéré, le

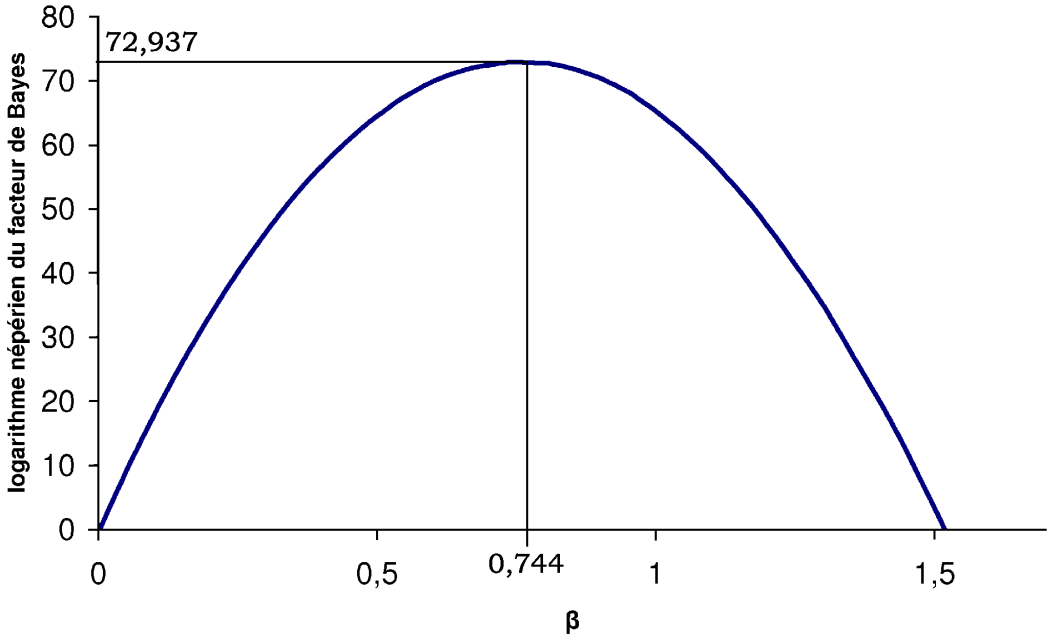


FIGURE 3.2 – Évolution du facteur de Bayes (en ordonnées) en fonction de β (en abscisses) pour le jeu de données GLOBIN (cf. chapitre 6).

modèle est optimal pour $\beta = 0,744$ et le logarithme népérien du facteur de Bayes vaut alors $72,937$.

3.6 Conclusion

Les différentes méthodes décrites dans ce chapitre sont utilisées et combinées à différents niveaux. Dans un premier article, nous avons défini un cadre statistique afin d'op-

timiser les paramètres d'un potentiel statistique à l'aide de la méthode du maximum de vraisemblance. Pour cela, nous avons fait appel à la méthode de descente de gradient. Toutefois, le gradient local est exprimé en fonction d'une espérance sur un ensemble de séquences suivant la distribution de probabilités $p(s|c,\theta)$ (cf. 4.2.3.3). Cette espérance est estimée à l'aide de séquences échantillonnées à l'aide de l'algorithme GS.

Cependant, l'échantillonnage de séquences à chaque pas de la descente de gradient est une étape très lourde d'un point de vue computationnel, et c'est pour cela que nous avons proposé une simplification de notre cadre méthodologique afin de ne plus avoir à combiner la descente de gradient et l'échantillonnage de séquences par MCMC (cf. chapitre 5).

Enfin, certaines des méthodes décrites ici sont utilisées au sein de l'application phylogénétique. Le cadre du modèle d'évolution est fourni par le programme développé par Nicolas Rodrigue [Rodrigue et al., 2005, Rodrigue et al., 2006, Rodrigue et al., 2007, Rodrigue et al., 2008b, Rodrigue et al., 2008a, Rodrigue et al., 2009, Rodrigue, 2007]. Ce programme utilise les potentiels optimisés par les méthodes décrites par les chapitres suivants (chapitres 4, 5, 6 et 7) et échantillonne les autres paramètres du modèle à partir de la distribution a posteriori à l'aide de l'algorithme MH. Notons également que ce programme peut utiliser l'algorithme GS lorsqu'il s'agit d'échantillonner les séquences ancestrales (comme celle présentée dans la figure 5.5), et cela en n'importe quel nœud de l'arbre. La dernière des méthodes présentées ici, le *path sampling*, est utilisée afin de calculer le facteur de Bayes et de comparer l'adéquation de modèles alternatifs dans le cadre phylogénétique.

Deuxième partie

Optimisation de potentiels pour le protein design et l'évolution moléculaire

Chapitre 4

Développement du cadre statistique

4.1 Introduction

Dans le chapitre 2, j'ai présenté quelques méthodes d'optimisation de potentiels statistiques couramment employés pour faire du *protein folding* ou du *protein design*. Certains de ces potentiels ont été intégrés dans des modèle d'évolution soumis à des contraintes structurales [Robinson et al., 2003, Bastolla et al., 2006]. Cependant, les méthodes que j'ai présentées dans le chapitre 2 pourraient ne pas être adaptées au contexte qui nous intéresse. Le problème se pose de savoir si les potentiels obtenus sous des approximations diverses (en particulier, l'approximation quasi-chimique) sont adaptés lorsqu'on les utilise dans un cadre SC. Plusieurs raisons peuvent être évoquées :

- la représentation simplifiée introduit des approximations. En effet, comme la structure n'est pas définie jusqu'à un niveau atomique, les paramètres du potentiel devront intégrer des valeurs moyennées sur tous les atomes composant les acides aminés.
- la forme du potentiel cause également des approximations. Par exemple, un potentiel de contact simple contient aussi des informations liées au potentiel d'accessibilité au solvant.
- un potentiel optimisé dans le cadre du *protein folding* est forcément optimisé à séquence constante, alors que le *protein design* implique une recherche dans l'espace des structures et dans l'espace des séquences. On peut donc se demander si un potentiel optimisé dans le contexte du *protein folding* serait à même de pouvoir décrire des variations au sein de la séquence d'intérêt.

Dans [Rodrigue et al., 2009], il a également été démontré que les potentiels optimisés pour le *protein folding* donnent de mauvais résultats lorsqu'on les incorpore tels quels dans un

modèle d'évolution soumis à des contraintes structurales.

Face à ces questions, le but avéré du travail présenté ici est de construire un cadre statistique général permettant de construire nos propres potentiels statistiques, afin de répondre aux besoins spécifiques des modèles structurellement contraints. Le formalisme statistique s'appuie sur le principe du maximum de vraisemblance : techniquement, on cherche à optimiser les potentiels de manière à prédire les séquences natives d'une série de structures tirées de la PDB. En utilisant des méthodes de validation croisée, on peut en outre tester de manière systématique les performances de plusieurs variations de la méthode, ou des formes alternatives de potentiel.

Les potentiels optimisés doivent avoir une forme simple, paramétrée, et facilement intégrable dans le modèle d'évolution. A cette fin, on a choisi de représenter la structure de la manière la plus grossière possible, c'est à dire en représentant chaque acide aminé par un seul pseudo-atome. Cette représentation a l'avantage de ne pas trop s'appesantir sur les chaînes latérales des acides aminés, et donc nous laisse une plus grande liberté dans le placement des acides aminés le long de la structure.

Il existe deux manières de présenter le problème qui nous intéresse. La première a été développée dans un contexte purement bayésien, alors que la seconde s'articule dans le contexte du modèle d'évolution [Choi et al., 2007]. Bien que la deuxième formulation, une fois posée, semble plus homogène avec le modèle d'évolution sous-jacent, nous avons commencé par poser le problème d'une manière bayésienne, en assignant des priors sur les séquences. La deuxième formulation sera présentée au chapitre 6. Ce chapitre correspond à notre article paru dans BMC Bioinformatics en 2006 [Kleinman et al., 2006].

4.2 A maximum likelihood framework for protein design

Authors : Claudia L. Kleinman, Nicolas Rodrigue, Cécile Bonnard, Nicolas Lartillot and Hervé Philippe.

4.2.1 Abstract

Background : The aim of protein design is to predict amino-acid sequences compatible with a given target structure. Traditionally envisioned as a purely thermodynamic question, this problem can also be understood in a wider context, where additional constraints are captured by learning the sequence patterns displayed by natural proteins of known conformation. In this latter perspective, however, we still need a theoretical formalization of the question, leading

to general and efficient learning methods, and allowing for the selection of fast and accurate objective functions quantifying sequence/structure compatibility.

Results : We propose a formulation of the protein design problem in terms of model-based statistical inference. Our framework uses the maximum likelihood principle to optimize the unknown parameters of a statistical potential, which we call an *inverse potential* to contrast with classical potentials used for structure prediction. We propose an implementation based on Markov chain Monte Carlo, in which the likelihood is maximized by gradient descent and is numerically estimated by thermodynamic integration. The fit of the models is evaluated by cross-validation. We apply this to a simple pairwise contact potential, supplemented with a solvent-accessibility term, and show that the resulting models have a better predictive power than currently available pairwise potentials. Furthermore, the model comparison method presented here allows one to measure the relative contribution of each component of the potential, and to choose the optimal number of accessibility classes, which turns out to be much higher than classically considered.

Conclusions : Altogether, this reformulation makes it possible to test a wide diversity of models, using different forms of potentials, or accounting for other factors than just the constraint of thermodynamic stability. Ultimately, such model-based statistical analyses may help to understand the forces shaping protein sequences, and driving their evolution.

4.2.2 Background

Predicting the sequences compatible with a given structure defines what is traditionally called the inverse folding problem, or more often, protein design [Drexler, 1981,Pabo, 1983, Ponders and Richards, 1987]. As suggested by the terminology, this question is usually considered in an engineering perspective : the aim is then to determine a sequence, or a set of sequences, that stably fold into a pre-specified conformation. In a thermodynamic perspective, this requirement translates into eliciting sequences that have lowest free energy under the target fold, compared to all possible alternative conformations. In principle, such a criterion would imply a search through the joint structure-sequence space, which is not feasible but for small on-lattice model proteins [Seno et al., 1996].

As an alternative to the engineering approach, a more evolutionary stance can be taken towards the inverse folding problem, in which case the aim would rather be to predict the sequences of *natural* proteins having the conformation of interest. Seen from this new point of view, the design problem raises new questions : natural proteins are the result of a complex evolutionary process, involving an intricate interplay between mutation and selection, and this probably entails many constraints directly related to the

native conformation, but nevertheless not equivalent to the mere requirement of structural stability. For instance, the requirement of fast and cooperative folding has an impact on the dispersion of contact energies [Abkevich et al., 1996]. For this and many other potential reasons, among all sequences predicted by classical engineering-oriented protein design, probably only a subset will look like natural proteins.

The evolutionary approach to protein design is particularly relevant to phylogenetic studies, where one of the current motivations is to develop the so-called structurally constrained models of protein evolution, i.e. models explicitly dependent on the protein's conformation, either for simulation purposes [Hellinga and Richards, 1994, Parisi and Echave, 2001, Bastolla et al., 2002, Bastolla et al., 2003], or in the context of phylogenetic inference [Robinson et al., 2003, Rodrigue et al., 2005]. In this framework, each substitution undergone by a protein during evolution has to be tested for its compatibility with the structure, in the context of the sequence that the protein displays at all other sites when the substitution occurs. Such repeated evaluation of the structure-sequence compatibility along a phylogenetic tree requires relevant and computationally very efficient scoring schemes/functions.

It is interesting to compare the different methods proposed thus far for performing protein design in light of this engineering/evolutionary distinction. A first direction of research has consisted in using all-atom semi-empirical force fields to evaluate the conformational free energy (reviewed in [Park et al., 2004]). These empirical methods have been applied to many theoretical and experimental cases, reaching a high level of accuracy. On the other hand, they are computationally heavy, mainly because of the side-chain positioning problem, and thus cannot be easily applied to structurally constrained phylogenetic models [Robinson et al., 2003, Rodrigue et al., 2005]. Concerns may also be expressed about their over-sensitivity to the native conformation, in particular in the core of the target structures and when the flexibility of the backbone is not accounted for [Wernisch et al., 2000, Larson et al., 2002]. But more importantly, approaches based on physical force fields are, by definition, exclusively focussed on the conformational stability, and thereby, completely oversee other potential factors shaping the sequences of biological proteins. As such, they are well suited for engineering synthetic proteins [Dahiyat et al., 1997], or for testing to what extent natural sequences are shaped by selection for protein stability [Jaramillo et al., 2002], but may not be sufficient for more general evolutionary purposes.

An alternative to the semi-empirical strategy consists in relying on knowledge-based, or statistical, potentials. These scoring functions mimic physical Boltzmann distributions, but merely encode statistical patterns present in the databases. Some of these potentials

were obtained under the quasi-chemical approximation, whereby frequencies of patterns, such as contacts between each pair of amino-acids, are transformed into energies using the Boltzmann law [Miyazawa and Jernigan, 1985, Sippl, 1993a, Godzik et al., 1995, Solis and Rackovsky, 2006]. Alternatively, contact energies can be obtained by maximizing the potential's predictive accuracy in a threading test [Hendlich et al., 1990, Maiorov and Crippen, 1992, Mirny and Shakhnovich, 1996, Bastolla et al., 2001]. In the present context, an advantage of these knowledge-based potentials, compared to semi-empirical force-fields, is that they should in principle capture all kinds of patterns that true biological sequences have, in relation to their conformation, and not only those directly related to thermodynamic stability. Furthermore, statistical potentials need not be defined at the atomic level, but can be based on a coarse-grained description of the protein's configuration, essentially by omitting the degrees of freedom associated to side chains. This allows faster computations, by avoiding the problem of searching through the rugged landscape of side-chain conformations. In addition, coarse-grained potentials could turn out to be an advantage, in that they will not recover the native sequence too faithfully. Most protein design procedures based on statistical potentials proposed until now have relied on coarse-grained, pairwise contact pseudo- nergies [Shakhnovich and Gutin, 1993, Kurosky and Deutsch, 1995, Deutsch and Kurowski, 1996, Seno et al., 1996, Seno et al., 1998, Micheletti et al., 1998, Banavar et al., 1998, Rossi et al., 2000, Rossi et al., 2001].

Yet, irrespective of the level of description adopted, currently available statistical potentials may not be ideal for protein design, since they have generally been optimized in the context of the folding problem, i.e. for maximizing the rate of correct structure prediction, given the sequence. In contrast, we would like to optimize the reciprocal prediction, namely, the sequences given the conformation. Several approaches have been proposed in this direction, consisting in maximizing the Z-score between the energy of the native sequence on the target conformation and its energy on a set of decoy sequences [Chiu and Goldstein, 1998b], or, alternatively, in applying a mean-square criterion on the values taken by the scoring function on each structure-sequence pair of the database [Seno et al., 1998]. However, these methods have thus far only been tested in cubic lattice protein models. In addition, they lack a firm theoretical basis. In particular, it would be interesting to guarantee optimal predictive power, and to have a robust methodology available to assess and compare the performance of alternative forms of statistical potentials.

Standard statistical theory provides such theoretical guarantees [Wald, 1949]. In the present case, the inverse folding problem can be formulated directly in terms of the probability of observing a sequence s given a conformation c, i.e. $p(s \mid c, \theta)$. This probability explicitly depends on the pre-specified model through a series of parameters, represented

here by θ. These may be, for instance, the coefficients of a pairwise potential, parameters describing compositional effects, secondary structure environment, solvent accessibility, etc. Taking the product over a database of P independent sequence- onformation pairs, $S = (s^p)_{p=1..P}$ and $C = (c^p)_{p=1..P}$, yields a joint probability

$$p(S \mid C, \theta) \quad = \quad \prod_p p(s^p \mid c^p, \theta) \tag{4.1}$$

which, as a function of θ, can be seen as a likelihood. The parameter θ is then learnt by maximizing the likelihood with respect to θ. Once this is done, sequences can be assessed, or sampled, under the optimal parameter value $\hat{\theta}$, by direct numerical evaluation of their probability, or by Monte Carlo sampling methods. Reformulated in this way, the method maximizes the predictive power of the potential, now in the structure-seeks-sequence direction. By construction, it yields the optimal parameter values that can be obtained for a given form of the potential. In addition, the fit of the model can be directly evaluated, based on the value of the likelihood obtained on a test data set, distinct from the learning set (cross-validation), giving a means of rigorous model selection. Finally, the statistical framework proposed here allows one to explicitly combine together, in a model dependent manner, all kinds of factors that we surmise may induce correlations between the structure and the sequence of proteins.

We have implemented this maximum likelihood (ML) procedure in a Markov chain Monte Carlo framework, and applied it to a simple case, using a contact potential, supplemented with a solvent accessibility term. Using cross-validation, we show that the resulting potentials yield a better fit than currently available potentials of the same form, and that combining solvent-accessibility considerations with contact energies is better than either alone. Furthermore, we find that solvent accessibility requires a more complex description than what is currently used. Ultimately, the overall method proposed in this work can be extended to a large spectrum of alternative models and statistical potentials.

4.2.3 Results

4.2.3.1 The probabilistic model

Let us consider a sequence $s = (s_i)_{i=1..N}$, of length N, and of conformation c. In its most general form, the method introduced here can work with any model M specifying the conditional probability of s given c, in terms of an unnormalized non negative function $q(s, c)$:

$$p(s \mid c, M) \quad = \quad \frac{q(s, c)}{\sum_s q(s, c)}. \tag{4.2}$$

To illustrate the method, we will apply it to a simple case, using a pairwise contact potential. The argument is as follows. First, by Bayes' theorem :

$$p(s \mid c, M) = \frac{p(c \mid s, M)\ p(s \mid M)}{\sum_s p(c \mid s, M)\ p(s \mid M)}. \tag{4.3}$$

If, in addition, we assume a uniform prior on s, we can simply relate equations 4.3 and 4.2 by posing $q(s, c) = p(c \mid s, M)$. Next, given a statistical potential $E(s, c)$, the conformational probability $p(c \mid s)$ can be expressed as a Boltzmann distribution :

$$p(c \mid s, M) = \frac{e^{-E(s,c)/kT}}{Z_s} \tag{4.4}$$

$$= e^{-(E(s,c)-F(s))/kT}, \tag{4.5}$$

where

$$Z_s = \sum_c e^{-E(s,c)/kT} \tag{4.6}$$

is a normalization constant, and

$$F(s) = -\ln Z_s. \tag{4.7}$$

T and k are the absolute temperature and the Boltzmann constant, respectively. Without loss of generality, it is possible to rescale the potential so that $kT = 1$, which we will do in the following.

Then, by defining the *inverse potential* :

$$G(s, c) = E(s, c) - F(s), \tag{4.8}$$

the conditional probability of sequence s reads as

$$p(s \mid c, \theta, M) = \frac{e^{-G(s,c)}}{Y}, \tag{4.9}$$

where

$$Y = \sum_{s'} e^{-G(s',c)} \tag{4.10}$$

is the normalization factor. Note that, contrary to the Z_s factor of equation 4.4, which was a sum over all conformations, the present factor Y is a sum over sequence space (all possible sequences of length N).

4.2.3.2 Statistical potentials

In the present work, we used a statistical potential made of two terms :

$$E(s,c) = \sum_{1\leq i<j\leq N} \Delta_{ij}\epsilon_{s_i s_j} + \sum_{1\leq i\leq N} \alpha_{s_i}^{v_i}. \tag{4.11}$$

The first term is a contact free energy : $\Delta_{ij} = 1$ if positions i and j are closer in space than a certain cut-off distance, and 0 otherwise, and ϵ_{ab} defines the contact energy between amino acids a and b. The second term encodes a solvent-accessibility free energy : for each position, α_a^d represents the free energy of amino acid a in the solvent accessibility class d, $a = 1..20$, and $d = 1..D$, where D is the total number of solvent accessibility classes considered.

Deriving the inverse potential requires the calculation of $F(s)$, which is already entirely specified by the potential E as a sum over all conformations. However, this computation is difficult in practice. As an alternative, we can give it a simple phenomenological form, inspired from the random energy model [Shakhnovich and Gutin, 1993, Sun et al., 1995, Seno et al., 1998] :

$$F(s) = -\sum_{1\leq i\leq N} \mu_{s_i}, \tag{4.12}$$

where the $(\mu_a)_{a=1..20}$ are unknown parameters, analogous to "chemical potentials" for the 20 amino acids.

Altogether, our parameter vector is made of three components : $\theta = (\alpha, \epsilon, \mu)$, and the inverse potential reads as :

$$G(s,c) = \sum_{1\leq i<j\leq N} \Delta_{ij}\epsilon_{s_i s_j} + \sum_{1\leq i\leq N} \alpha_{s_i}^{v_i} + \sum_{1\leq i\leq N} \mu_{s_i}. \tag{4.13}$$

Note that the probability defined by equation 4.9 is invariant under the following transformation :

$$\mu'_a = \mu_a + J_1, \tag{4.14}$$
$$\epsilon'_{ab} = \epsilon_{ab} + J_2, \tag{4.15}$$
$$\alpha'^d_a = \alpha^d_a + J_3, \tag{4.16}$$

where J_1, J_2 and J_3 are arbitrary real constants. Therefore, to ensure identifiability of our

probabilistic model, we enforce the following constraints :

$$\sum_a \mu_a = 0, \tag{4.17}$$

$$\sum_{ab} \epsilon_{ab} = 0, \tag{4.18}$$

$$\sum_a \alpha_a^d = 0\,,\ d = 1..D. \tag{4.19}$$

A series of alternative inverse potentials can be obtained by suppressing the first or the second of the components of equation 4.13. In the present work, we tested the following combinations :

- μ,
- $\alpha + \mu$,
- $\epsilon + \mu$,
- $\epsilon + \alpha + \mu$.

We also explored various numbers of accessibility classes, with D ranging from 2 to 20. Alternatively, the ϵ component can be fixed to values of a contact potential obtained by other authors (MJ) [Miyazawa and Jernigan, 1985]. In this case, we must add a multiplicative scaling factor λ in front of the contact component to account for the fact that these potentials are normalized differently :

$$G(s,c) = \lambda \sum_{1\le i<j\le N} \Delta_{ij}\epsilon^{MJ}_{s_i s_j} + \sum_{1\le i\le N} \mu_{s_i}. \tag{4.20}$$

The scaling factor is optimized by ML, along with μ.

4.2.3.3 Optimizing the potentials by gradient descent

If we now consider a database, made of P protein sequences $S = (s^p)_{p=1..P}$, of respective lengths N_p and their corresponding three dimensional structures $C = (c^p)_{p=1..P}$, the probability of observing the whole database, which we define as the *likelihood* $L(\theta)$, is the product of the probabilities of observing each protein independently :

$$L(\theta) = p(S \mid C, \theta) \tag{4.21}$$

$$= \prod_p p(s^p \mid c^p, \theta) \tag{4.22}$$

$$= \frac{e^{-G(S,C)}}{Y} \tag{4.23}$$

where

$$G(S,C) = \sum_p G(s^p, c^p) \tag{4.24}$$

is the inverse potential summed over the database, and

$$Y = \sum_{S'} e^{-G(S',C)} \tag{4.25}$$

is the corresponding normalization constant. Since it is more convenient to work on minus the logarithm of the probability, we define the score ω :

$$\begin{aligned} \omega(\theta) &= -\ln L(\theta) && (4.26)\\ &= G(S,C) + \ln Y. && (4.27) \end{aligned}$$

We wish to maximize the likelihood, or equivalently, minimize ω, with respect to θ. We do this by gradient descent, based on a numerical evaluation of the derivative of ω (see methods). The overall method is akin to an Expectation Maximization algorithm [Dempster et al., 1977]. In fact, it can be seen as a differential version of Dempster's method, and therefore, we call it *differential EM*.

The derivative of ω reads as :

$$\frac{\partial \omega}{\partial \theta} = \frac{\partial G(S,C)}{\partial \theta} + \frac{\partial \ln Y}{\partial \theta}. \tag{4.28}$$

Applying the partition function formalism to equation 4.25, we can express the second term as an expectation over $p(S' \mid C, \theta)$:

$$\begin{aligned} \frac{\partial \ln Y}{\partial \theta} &= \frac{1}{Y}\frac{\partial Y}{\partial \theta} && (4.29)\\ &= -\frac{1}{Y}\sum_{S'} \frac{\partial G(S',C)}{\partial \theta} e^{-G(S',C)} && (4.30)\\ &= -\sum_{S'} \frac{\partial G(S',C)}{\partial \theta} p(S' \mid C, \theta) && (4.31)\\ &= -\langle \frac{\partial G}{\partial \theta} \rangle && (4.32) \end{aligned}$$

which leads us to the following expression for the derivative of ω :

$$\frac{\partial \omega}{\partial \theta} = \frac{\partial G(S,C)}{\partial \theta} - \langle \frac{\partial G}{\partial \theta} \rangle. \tag{4.33}$$

The computation of the first term in this equation is straightforward, while the second term must be estimated numerically. In order to do so, we obtain a sample $(S_h)_{h=1..K_{EM}}$ drawn from $p(S \mid C, \theta)$ by a Gibbs sampling algorithm similar to that of Robinson et al. [Robinson et al., 2003] (see methods).

Applying formula 4.33 on the inverse potential 4.13 yields the following expressions for the derivatives :

$$\frac{\partial \omega}{\partial \epsilon_{ab}} = -\left[n_{ab} - \langle n_{ab} \rangle\right], \tag{4.34}$$

where n_{ab} is the number of contacts between amino acids a and b observed in the database, and $\langle n_{ab} \rangle$ is its expectation over the probability distribution $p(S' \mid C, \theta)$. Formula 4.34 thus leads to an intuitive characterization of the maximum likelihood estimate $\hat{\epsilon}$: it is the value of ϵ such that the average number of each type of contact predicted by the potential matches the number observed in the database. Following a similar derivation :

$$\frac{\partial \omega}{\partial \mu_a} = -\left[m_a - \langle m_a \rangle\right], \tag{4.35}$$

where m_a is the total number of amino acids of type a, and

$$\frac{\partial \omega}{\partial \alpha_a^d} = -\left[l_a^d - \langle l_a^d \rangle\right], \tag{4.36}$$

where l_a^d is the total number of amino acids of type a belonging to solvent-accessibility class d.

We first performed an optimization of the pure contact potential ($\epsilon + \mu$-potential) on each data set. Figure 4.1 shows the evolution of the scoring function ω and of the contact potential during the gradient descent. As can be seen from these traceplots, the differential EM algorithm converges after a few hundred cycles. The scoring function stabilizes at around 272,000 natural units of logarithm (nits), and then fluctuates by up to 25 nits around this value. These fluctuations are mainly due to the finite size of the sample of sequences on which the derivative of $\ln Y$ is evaluated and, to a lesser extent, to the error on the estimation of $\ln Y$ by thermodynamic integration. In any case, these errors are small compared to the differences between scores obtained with alternative models (see below).

The evolution of the potential for some residue pairs is shown in figure 4.1.b and 4.1.c. Effects in the final values due to residue polarity are easily seen : known favorable interactions such as glutamate-lysine or the hydrophobic isoleucine-valine have a lower contact energy, while known unfavorable interactions, such as glutamate-glutamate, have higher energies, indicating that the potentials obtained are biologically reasonable.

The potentials obtained in two independent runs are virtually identical (figure 4.2.a), indicating that the gradient descent does not get trapped into local minima. We can also compare the values of the potential for two distinct data sets of equivalent size, DS1 and DS2 (figure 4.2.b), which uncovers a greater discrepancy than for two independent runs

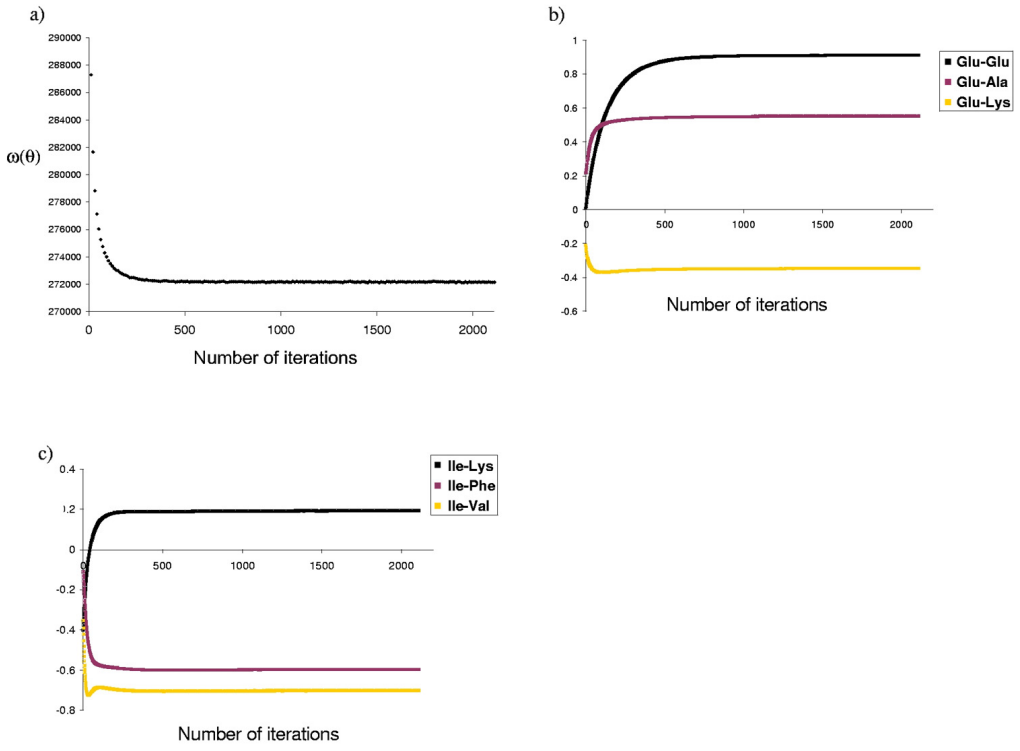


FIGURE 4.1 – Convergence of the optimization procedure - Traceplots illustrating the convergence of the differential EM method in the optimization of contact potentials, on data set DS1. Are shown, as a function of the number of iterations **(a)** the score $\omega(\theta) = -\ln p(S \mid C, \theta)$, **(b)** nd **(c)** examples of pairwise contact energies obtained for some amino acid pairs.

on the same data set DS1. The correlation is high, however, suggesting that data sets are large enough for the learning procedure to reach stability. In addition, these differences are small compared to the discrepancy between the potential obtained by our method and that of Miyazawa & Jernigan (figure 4.2.c).

4.2.3.4 Model comparison

The same optimization procedure was applied to the potential consisting only of the solvent accessibility term $(\alpha + \mu)$, with an increasing number of accessibility classes, and to the combined $(\epsilon + \alpha + \mu)$ potential. The resulting log likelihood scores cannot directly be compared, since the models do not have the same dimensionality. We therefore applied a 2-fold cross-validation procedure (CV), consisting in learning the potential on DS2, and testing it on DS1, and vice versa.

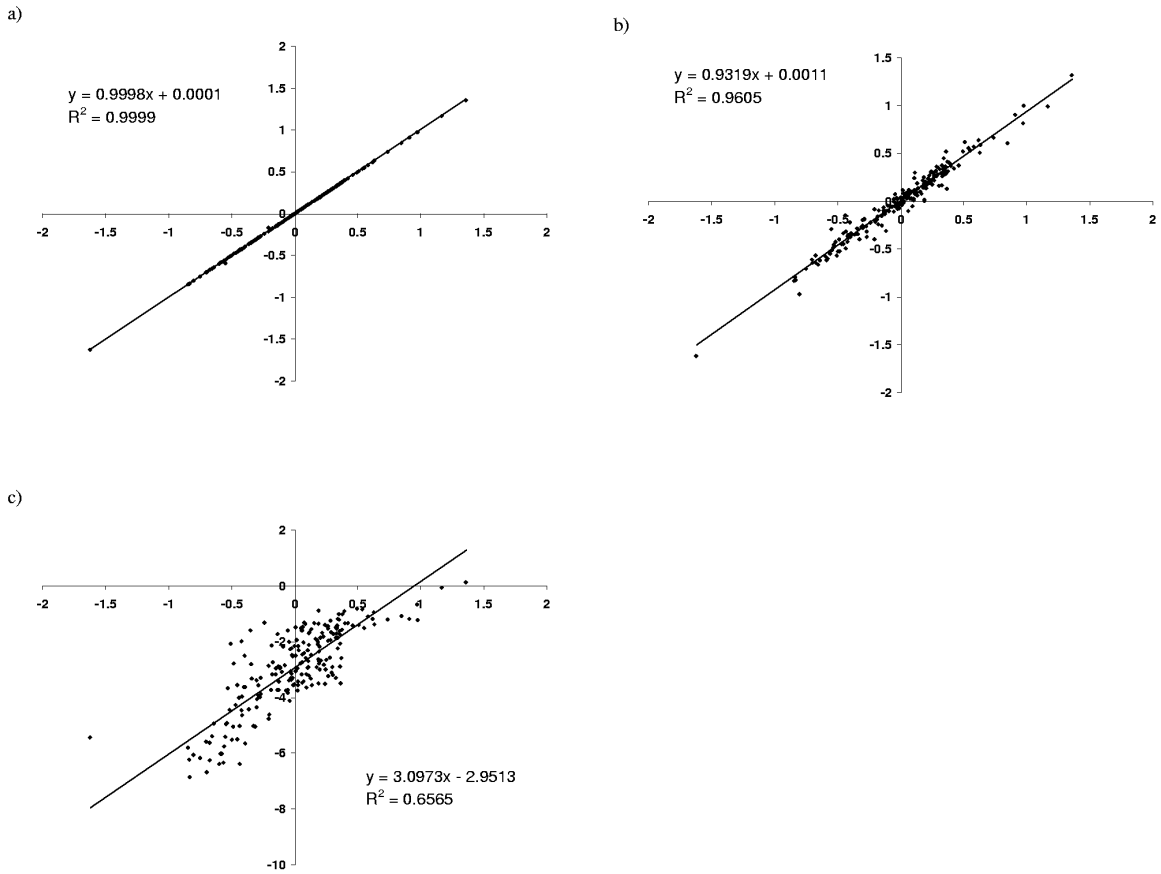


FIGURE 4.2 – XY-comparisons of pairwise contact potentials. **(a)** two independent runs on the same data set DS1, **(b)** two runs, on data sets DS1 (X-axis) and DS2 (Y-axis); **(c)** Miyazawa and Jernigan's potential, compared to that obtained on DS1.

The evolution of the CV score as a function of the number of accessibility classes (D) is shown in figure 4.3. When D increases, the fit of the model improves, until a point is reached where the penalization for model dimensionality starts to dominate the score. The optimal number of classes obtained is 14 to 16, depending on the form of the potential studied, although 4 to 6 classes is sufficient to attain 90% of the fit improvement.

The scores obtained for the different models tested are reported in figure 4.4. We also included in the comparison the Miyazawa and Jernigan potential [Miyazawa and Jernigan, 1985]. The contact potential performs better than the pure solvent accessibility potential, and the combination of both terms is the most informative. Miyazawa and Jernigan's potential results in a poorer fit improvement than any of the other models.

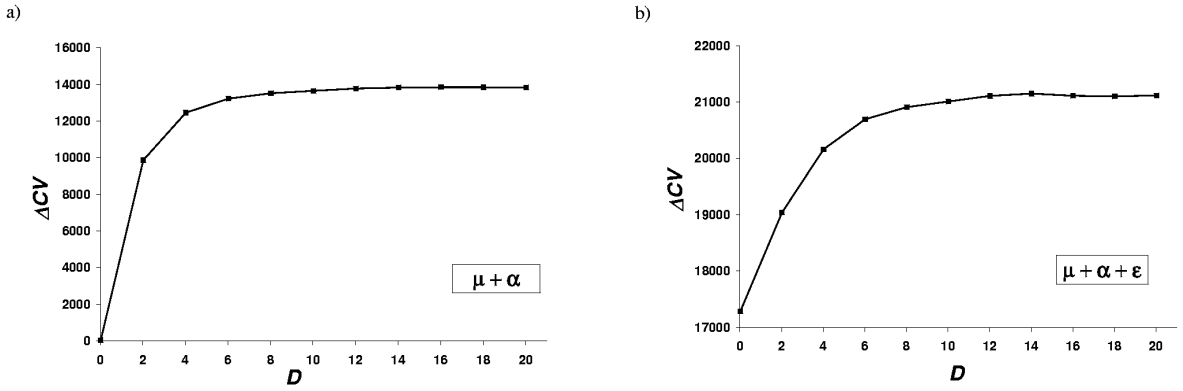


FIGURE 4.3 – Effect of the solvent accessibility definition on the potential. Gain in cross-validation score (see Methods) as a function of the number of accessibility classes. The average gain for the 2-fold cross-validation experiment is shown. **(a)** Inverse potential consisting in solvent accessibility terms only, and **(b)** inverse potential combining contact and solvent accessibility terms.

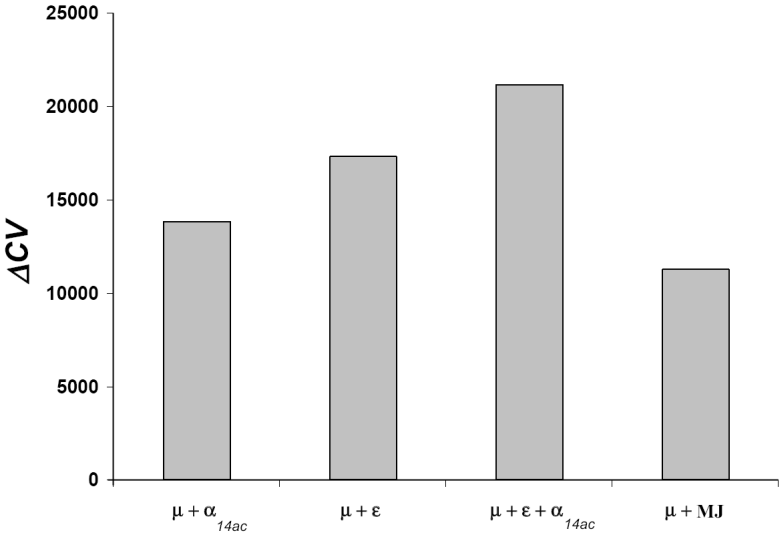


FIGURE 4.4 – Model comparison. Cross-validation (CV) scores obtained for the different forms of potentials tested. The average gain (relative to the CV score obtained with the flat potential μ, see Methods) for the 2-fold cross-validation experiment is reported. α_{14ac} : solvent accessibility potential, 14 accessibility classes ; ϵ : contact potential ; MJ : Miyazawa and Jernigan's potential.

4.2.3.5 Specificity of the designed sequences

Once an optimal value of θ is obtained, properties of the sequences induced by the models can be investigated by sampling sequences from $p(s \mid c, \theta)$, using this optimal value of θ. In particular, we tested to what extent the sequences proposed by our method met the requirement of specificity, i.e. the condition that the sequences designed on a given conformation c indeed have c as their unique ground state. More precisely, we generated 20 sequences by Gibbs sampling for 60 randomly chosen structures [see Additional file 8], i.e. 1,200 sequences for each potential, and performed a fold recognition experiment for the designed sequences, monitoring the score for the target fold using THREADER [Jones et al., 1992c] (figure 4.5 and Table 4.1).

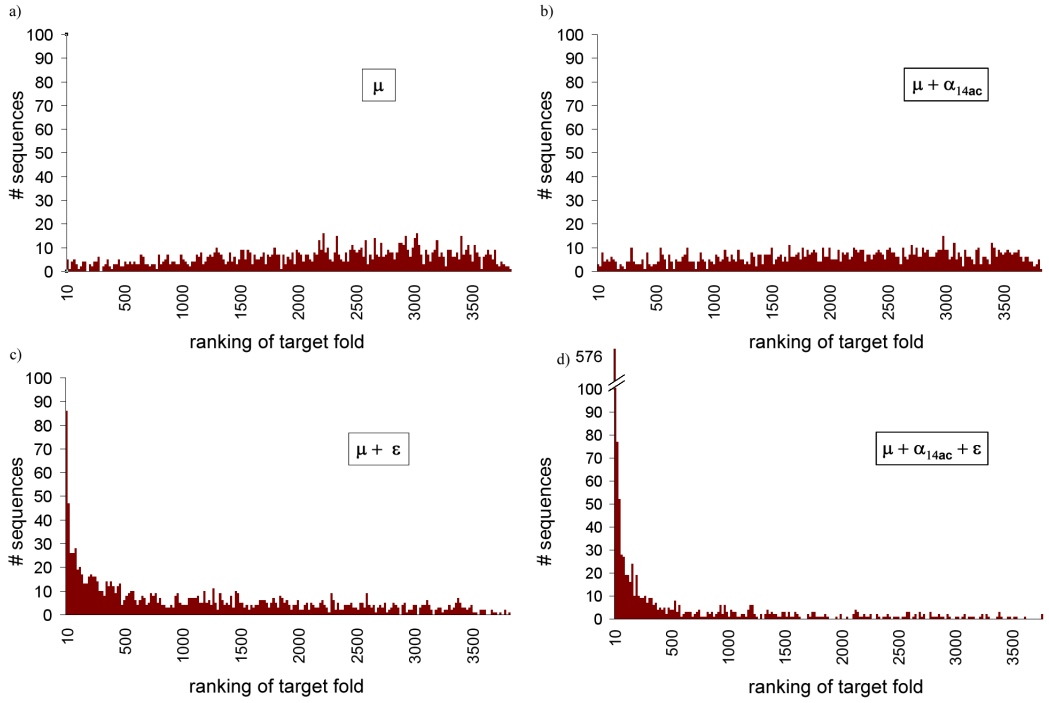


FIGURE 4.5 – Design specificity. Histograms of the ranking of the target structure in a fold recognition experiment using THREADER. 20 sequences were generate for 60 randomly chosen structures, using (a) a flat (μ) potential, (b) a solvent accessibility, 14 classes ($\mu + \alpha_{14ac}$) potential, (c) a contact ($\mu + \epsilon$) potential, and (d) the combined ($\mu + \alpha_{14ac} + \epsilon$) potential.

The solvent accessibility potential alone ($\alpha_{14ac} + \mu$, figure 4.5b) is not sufficient to provide specificity to the designed sequences, and behaves almost as poorly as the flat potential (μ, figure 4.5a). A mild improvement is seen when using the contact potential

Potential	Average Z-score ratio	SDev Z-score ratio	Ranking (median)	Target fold in top 1% (A)	Target fold in top 10%	Average seq. identity	Correlation between (A) and mean entropy/site
μ	-0.12	0.18	2249	0.5%	4.8%	5.76 %	-0.26
$\mu + \alpha_{14ac}$	-0.10	0.18	2090	0.4%	6.3%	6.65 %	-0.04
$\mu + \epsilon$	0.13	0.16	816.8	10.7%	33.5%	6.69 %	0.23
$\mu + \alpha_{14ac} + \epsilon$	0.45	0.23	32.7	53.6%	77.5%	7.82 %	0.64

TABLE 4.1 – Specificity of designed sequences - Scores of a fold recognition experiment for designed sequences (see Methods). 1,200 sequences were sampled from $p(s \mid c, \theta)$ for each potential, and submitted to THREADER for fold recognition. Z-score ratio : Z-score of designed sequence / Z-score of native sequence in target fold.

($\epsilon + \mu$, figure 4.5c) : for 10% of the designed sequences the target fold is found among the best scoring folds (Table 4.1), and the distribution of this ranking is skewed towards lower values. However, it is only with the combined potential ($\epsilon + \alpha_{14ac} + \mu$, figure 4.5d) that a significant improvement is observed : for more than half of the designed sequences the target fold is found among the best 1% scoring folds, even though the average sequence identity with the native sequence is less than 10% in all cases (Table 4.1).

We also tested a subset of 120 randomly chosen designed sequences using another fold recognition program, LOOPP [Meller and Elber, 2001]. LOOPP is based on a combination of several structure prediction methods, based on threading, secondary structure, sequence profile and exposed surface area prediction. The results obtained with this program were similar to those of THREADER : for 51.2% of the designed sequences using the combined ($\epsilon + \alpha_{14ac} + \mu$) potential, the target fold was found as the first hit, and for 67.2% the target fold was found among the first 10 hits.

In contrast, many of the current fold recognition programs based on sequence profile methods produced no significant hits (data not shown), which is not surprising, given that our sampling algorithm produces highly divergent sequences, with no similarity to any natural protein.

4.2.4 Discussion

The central idea of the present work is to reformulate the problem of devising statistical potentials for protein design as a statistical inference problem. This reformulation, based on the maximum likelihood (ML) principle, led us naturally to a gradient descent method, with the only additional aspect being that the gradient to follow is itself estimated by Monte-Carlo averaging.

The main advantage of this ML framework is that it guarantees an optimal predictive power of the resulting potential. In addition, it is very general, and can in principle be applied to any form of statistical potential. In particular, it is not restricted to coarse grained descriptions of proteins, and it could also be applied at the atomic level.

Interestingly, our gradient descent method turns out to be similar in spirit to an iterative scheme proposed by Thomas and Dill [Thomas and Dill, 1996a], although in that case the purpose was to optimize a potential in the context of the folding problem. Specifically, Thomas and Dill tune the potential so as to match the observed and expected number of contacts of each type, except that their expectation is taken on a set of alternative conformations, for a fixed sequence, whereas we take the expectation on a set of alternative sequences, on the conformation of interest. Note that Thomas and Dill derived

their method from intuitive arguments, and not as a mathematical consequence of the ML principle.

These two alternative optimization schemes, obtained by normalizing either over the sequence or over the structure space, are quite distinct, at least conceptually. How the resulting potentials would differ in practice is more difficult to evaluate. Among other things, it will depend on how the approximation of lnZ_s based on the random energy model works. In the eventuality that it does not work well, it is likely that the contact term of our inverse potential will in fact combine two things : the information corresponding to the conformational energy of the sequence itself, which is also encoded in classical potentials optimized for threading, plus some information coming from the decoy term $\ln Z_s$. A way to settle this question would be to optimize a contact potential using, on the same learning set, both normalization schemes, and then compare the resulting values as well as their predictive powers.

4.2.4.1 Model assessment and comparison

The methodological framework proposed here offers reliable criteria for comparing the empirical fit of alternative models on real data. In this respect, it should be noted that the lack of a reliable objective criterion for evaluating different statistical potentials has often been invoked for justifying the use of on-lattice idealized models [Mirny and Shakhnovich, 1996]. However, on-lattice approaches are only moderately interesting, as they completely ignore the problem of the robustness of the learning method to model violation. Coarse-grained statistical potentials are by definition over-simplified models of proteins, and therefore, model violation is an intrinsic feature of the protein design problem. In this respect, the statistical language is interesting, since it is still valid, even for fitting and assessing models that are known to be imperfect.

On the other hand, the intuitive idea underlying cross-validation, i.e. measuring the rate of prediction of the native sequence, is quite simple, and has been invoked and used several times previously [Sun et al., 1995, Micheletti et al., 1998, Kono and Saven, 2001, Rossi et al., 2001, Jaramillo et al., 2002]. What we propose here is a better formalization of this idea. Note that in contrast to previous methods, we do not measure the *marginal* native prediction rate at each site, but the *joint* probability of the native sequence. This can be important, as it accounts for possible correlations in the predictive distribution. For instance, two given positions may not display any particular pattern, when considered marginally, but may jointly follow charge or steric compensatory patterns. These phenomena will not be taken into account in the overall fit of the potential when measuring the

marginal prediction rate, as is usually done. Technically speaking, the joint probability of the native sequence on the corresponding structure is extremely small, and cannot be evaluated just by counting the frequency at which the native sequence appears in the sample obtained by Gibbs sampling. For this, more elaborate numerical methods, such as thermodynamic integration, are required.

In the present case, the comparison between alternative models has allowed us to measure the relative contribution of each term of the potential and to refine the protein representation. The contact component turns out to be the most informative (figure 4.4), although it should be complemented with other energetic forms. Here, we have tested the addition of a solvent accessibility component, which significantly improves the fit of the model. Contact information and solvent exposure are correlated, which is reflected in the fact that the fit improvement of each term is not additive.

Our model comparison method also gives us a direct way of choosing the optimal number of solvent accessibility classes (figure 4.3). Here, we found a number of 14 to 16 classes, which is higher than what one may have expected and than what is usually used. Note that this number depends on the way the classes are defined ; here, the classes are based on quantiles, but as an alternative, we also tried a linear definition (evenly splitting the whole range of accessibility surfaces into D bins), which gave us an even higher optimal number of classes (20 classes, data not shown). In general, the present methodology could be used to investigate different definitions of accessibility classes, to refine the pairwise contact definition, or any other elements of the structure representation included in the potential.

The fact that our potential has a significantly better predictive power than that of Miyazawa and Jernigan (MJ, figure 4.4) is trivially expected, by construction of the ML potential. What is more surprising is that the MJ matrix is less fit than a simple solvent-accessibility profile. A possible explanation would be that Miyazawa and Jernigan's potential is based on the quasi-chemical approximation, which is now known to be somewhat drastic [Godzik et al., 1995, Thomas and Dill, 1996b, Skolnick et al., 1997], as it neglects correlations between observed pairing frequencies, due to chain connectivity and multiple contacts. Alternatively, it could mean that potentials optimized for folding are really not suited for protein design purposes. Testing other pairwise contact potentials, in particular those that do not rely on the quasi-chemical approximation [Tiana et al., 2004, Bastolla et al., 2001, Maiorov and Crippen, 1992, Tobi and Elber, 2000, Vendruscolo et al., 2000], would be a way to address this issue.

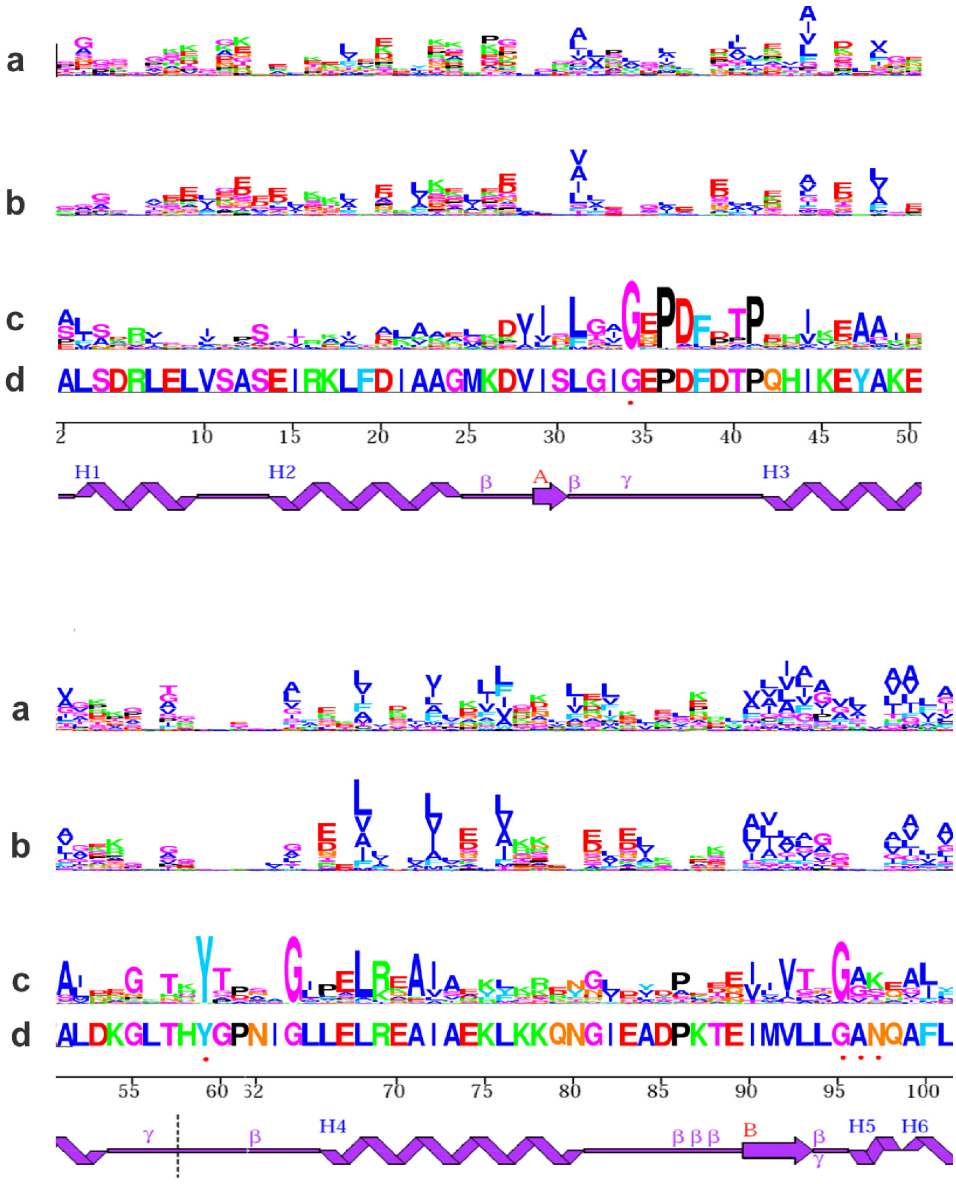


FIGURE 4.6 – Site-specific profiles. Sequence logos of site-specific profiles induced on an alpha-aminotransferase ([PDB :1GDE], chain A), using a contact + solvent accessibility (14 classes) potential. From top to bottom : **(a)** marginal profiles, **(b)** leave-one-out profiles, **(c)** empirical profiles from a multiple sequence alignment of 162 sequences [see Additional file 4], and **(d)** native sequence of the reference protein. Secondary structure representation was taken from PDBsum [Laskowski et al., 2005]. Red dot : residue interaction with ligand. Only the first 100 amino acids are shown ; sequence logos for the whole protein are available as supplementary material [see Additional file 5][see Additional file 6].

4.2.4.2 Sequence sampling

The method that we propose in this work is probabilistic in essence. As such, it offers a very natural framework for investigating the patterns induced by the models on distributions of sequences.

Specificity of the designed sequences A sequence s designed for a target conformation c should not only be compatible with c, but also incompatible with competing folds. A rigorous solution to this problem involves a simultaneous search over the sequence and conformation space. It is possible, however, to achieve specificity without explicitly seeking to penalize competing states (*negative* design), if we rely on the approximation based on the random energy model, where the normalization constant of equation 4.4 can be considered as a function of the sequence composition only [Shakhnovich and Gutin, 1993, Koehl and Levitt, 1999a]. In our case, the normalization of the likelihood will also play an important role : since the total probability over all possible sequences has to be 1, maximizing the probability for a given sequence s_1 on its native conformation c_1 will lower the probability that another natural sequence s_2, with native conformation c_2, also gets a high probability on c_1. When many sequences are learnt in parallel, this phenomenon should ultimately favor specificity of s_2 on c_2, compared to all other conformations of the data set.

On the other hand, the extent to which the specificity is achieved will depend on the actual form of the potential used, as well as on the data base used for learning. To address this question, we produced a large number of sequences with four different potentials, and checked their ability to recognize the target fold, as measured by the Z-score ratio or by the ranking of the target structure in a fold recognition experiment. Indeed, an improvement of specificity is observed when using better potentials, suggesting that the method is effectively capturing specific dependencies between the conformation and the sequence of the proteins in the learning set, even for the simple forms of potentials tested here. For the combined $(\epsilon + \alpha_{14ac} + \mu)$ potential, the average Z-score ratio of the designed sequences is similar to what has been reported for other protein design algorithms [Koehl and Levitt, 1999a]. Conversely, this also suggests that a more sophisticated potential may further improve the specificity of the sequences designed using our algorithm.

Conformation-dependent site-specific profiles To compare natural protein sequences with those predicted by the optimized potentials, marginal, leave-one-out and empirical profiles (see methods) were generated for the 60 proteins used in the design specificity experiment described above ; the profiles obtained for the best and the worst scoring

structures are provided as supplementary materials [see Additional file 7]. Overall, leave-one-out profiles (figure 4.6.a) and marginal profiles (figure 4.6b) do not display significant differences in the discriminative power between sites : the mean Shannon entropy per site is 0.743 ± 0.366 for marginal profiles, and 0.696 ± 0.428 for leave-one-out profiles. It is worth noting that the mean entropy per site for each protein, and the corresponding standard deviation, i.e. the average amount of information at each site and the variation between sites, are both correlated with the performance of the particular protein in the fold recognition experiment, and this, only for the combined ($\epsilon + \alpha_{14ac} + \mu$) potential (Table 4.1).

A detailed analysis of the leave-one-out profiles for a particular case, an alpha-amino-transferase, may be useful to understand which type of information is effectively captured by the potential, and which is not captured at all, thereby suggesting possible ways of improving the current form of potential.

First, regions of the protein that show little secondary structure (such as in positions 32-40, 55-65 and 82-88) contain less information (mean entropy per site = 0.756) than regions with local structure (mean entropy per site = 0.856). This is not surprising, since these regions typically have fewer contacts between residues, and thus the amount of information included in the protein representation is lower.

Concerning regions with defined secondary structure, residue polarity is the information most easily captured. Charged residues are also distinctively inferred, as well as glycines, to a lesser extent (e.g. glycine 64, 81 and 95 – the latter predicted at position 94 or 95). In contrast, prolines are rarely correctly predicted, which is expected, since the properties most distinctive of prolines (such as phi-psi dihedral angles or local secondary structure) are not included in this particular form of potential.

Interestingly, some residues that have a crucial importance for the protein structure or function fail to be predicted, simply because the properties conferring their importance are not included in the protein description. This is the case of the amino acids that are in close interaction with a ligand (positions 34, 59, 96, 97).

Finally, the leave-one-out profiles display an interesting behavior with respect to positions where the amino-acid present in the reference sequence is not at all conserved in other members of the family. In some cases, they simply do not predict anything (e.g. glycines 24 and 60, or leucine 9, isoleucine 21, and alanine 23), which suggests that their limited importance in structure stability or function is recognized by the inverse potential. In other cases, the natural profile is even reproduced in the leave-one-out profile, instead of the amino acid of the reference sequence ; such is the case for phenylalanine 100.

4.2.5 Conclusions

As illustrated by the sequence logos and the fold recognition experiments performed above, the predictive power of the models proposed here is encouraging, but nevertheless still weak. It is not yet clear to what extent this is due to the specific choice made concerning the form of the statistical potential, to the approximation of $\ln Z_s$ as a function of the sole composition of the sequence, or to yet other reasons. Most probably, we are facing a combination of several factors. The methods proposed here can now be used to address these difficult questions empirically.

In one direction, other approximations of $\ln Z_s$, less drastic than the random energy model, but still accessible in practice, can be investigated. For instance, following Deutsch and Kurozky (1996), the conditional probability of a sequence could be defined as :

$$p(s \mid c) \quad \propto \quad e^{-[E(S,C)-\langle E(S)\rangle]} p(s) \tag{4.37}$$

where the expectation $\langle \cdot \rangle$ is taken over a pre-defined set of decoy conformations. More sophisticated Monte Carlo methods, jointly sampling the sequence and conformation spaces, can also be imagined, in order to get more precise evaluations of $\ln Z_s$, while staying in the same global maximum likelihood formalism.

On the other hand, all the many statistical potentials that have been proposed over the last fifteen years may in principle be investigated in the same way as we have done here. In particular, distance-dependent potentials [Sippl, 1990] and main-chain dihedral angle potentials [Betancourt and Skolnik, 2004], which imply a richer representation of the protein structure, may result in models of greater predictive power. Other ways of implicitly considering side-chain conformation may also be easily incorporated into the model.

In a completely different perspective, it is possible to devise probabilistic models that are not exclusively defined in terms of a conformational free energy, even in a formal way. For instance, additional terms, concerning secondary structure aspects, interactions between successive positions along the sequence, or terms related to the folding constraints, can all be combined in an additive manner in the inverse potential. In fact, the model need not even be formulated in terms of a Boltzmann distribution, as long as the parameters are fitted by ML, and the predictive power of the resulting models is evaluated in a systematic way. Altogether, this amounts to setting up a robust statistical framework helping us to understand how, and to what extent, the sequences of natural proteins are determined by protein structure.

4.2.6 Methods

4.2.6.1 Structure representation

We used Miyazawa and Jernigan's definition of contacts [Miyazawa and Jernigan, 1985] : each residue is represented by the center of its side chain atom positions ; the positions of C^{α} atoms are used for glycine. Residues whose centers are closer than 6.5Å are defined to be in contact. The accessible surface of a residue is defined as the atomic accessible area when a probe of the radius of a molecule of water is rolled around the Van der Waal's surface of the protein [Lee and Richards, 1971]. We used the program Naccess [Hubbard and Thornton, 1993] to make this calculation. When treating PDB files with multiple chains, solvent accessibility was calculated taking into account all molecules in the structure. The accessibility classes (percentage relative to the accessibility in Ala-X-Ala fully extended tripeptide) were defined so as to generate D equal-sized subsets of sites. The complete definition of accessibility classes is available as supporting material[see Additional file 1].

4.2.6.2 Monte Carlo implementation

In order to calculate the derivative of ω in the gradient descent procedure, expectations with respect to $p(S' \mid C, \theta)$ in equation 4.33 are evaluated numerically. A sample $(S_h)_{h=1..K_{EM}}$ drawn from $p(S \mid C, \theta)$ is obtained by a Gibbs sampling algorithm similar to that of Robinson et al. [Robinson et al., 2003]. The elementary cycle of our Gibbs sampler is as follows : for each $p = 1..P$, and for each $i = 1..N_p$, each of the 20 amino acids is proposed at site i of protein p, by successively setting $s_i^p = a$, for all $a = 1..20$; in each case, the energy change ΔG_a induced by this point substitution is evaluated ; then, s_i^p is set to amino acid a with probability $p_a \propto e^{-\Delta G_a}$. After Q cycles of burnin, a series of $h = 1..K_{EM}$ cycles are performed, and after each cycle, the current sequence, S_h, is recorded. Once the sample is obtained, the expectation (4.32) is evaluated as

$$\langle \frac{\partial G}{\partial \theta} \rangle \simeq \frac{1}{K_{EM}} \sum_{h=1}^{K_{EM}} \frac{\partial G(S_h, C)}{\partial \theta} \tag{4.38}$$

and the derivative of ω with respect to θ follows immediately.

The overall gradient descent procedure runs as follows : we start from a random potential θ_0 and a random set of sequences, and perform the following iterative scheme :

– perform Q Gibbs cycles for the burnin, and K_{EM} additional cycles for the sampling itself. Keep the final sequences as the starting point of the next cycle.

– update θ by gradient descent, based on the estimate of the gradient obtained over the sample :

$$\theta_{n+1} = \theta_n - \delta\theta . \frac{\partial \omega(S)}{\partial \theta} \tag{4.39}$$

where . is a scalar product, and $\delta\theta$ is a step-vector. In practice, the coefficients of $\delta\theta$ are tuned empirically, allowing three degrees of freedom, for the α, the ϵ, and the μ component of the potential respectively.

– iterate.

As a stopping rule, we monitor the evolution of $\omega(\theta)$ itself, which we evaluate every 100 steps by a numerical procedure (see below), and stop when $\omega(\theta)$ has stabilized. In practice, we used $Q = 100$ and $K_{EM} = 100$. At first sight, it would seem that a larger number of points K_{EM} would be needed to get a precise expectation, but in the present case one can rely on the self-averaging of the derivatives across the 100,000 sites of the database.

4.2.6.3 Likelihood evaluation

The difficult part in estimating the likelihood (or equivalently $\omega(\theta)$), for a given value of θ, is to obtain an evaluation of $\ln Y$. We do this by thermodynamic integration, or path sampling [Ogata, 1989, Gelman, 1998], using the quasi-static method which we developed previously [Lartillot and Philippe, 2006].

First, for $0 \leq \beta \leq 1$, we define

$$G_\beta(s,c) = \beta \left(\sum_{1 \leq i < j \leq N} \Delta_{ij} \epsilon_{s_i s_j} + \sum_{1 \leq i \leq N} \alpha_{s_i}^{v_i} \right) + \sum_{1 \leq i \leq N} \mu_{s_i}. \tag{4.40}$$

The associated probability distribution is :

$$p_\beta(s \mid c, \theta) = \frac{e^{-G_\beta(s,c)}}{Y_\beta}, \tag{4.41}$$

$$Y_\beta = \sum_{s'} e^{-G_\beta(s',c)}. \tag{4.42}$$

What we are looking for is $\ln Y_1$. As for $\ln Y_0$, it factors out, and can be computed directly :

$$\ln Y_0 = N \ln \left(\sum_{a=1}^{20} e^{-\mu_a} \right). \tag{4.43}$$

We can thus equivalently evaluate the difference $\ln Y_1 - \ln Y_0$. To do this, we rely on the following identity :

$$\ln Y_1 - \ln Y_0 = \int_0^1 \frac{\partial \ln Y}{\partial \beta} d\beta \tag{4.44}$$

$$= \int_0^1 \langle \frac{\partial G}{\partial \beta} \rangle_\beta \, d\beta, \tag{4.45}$$

where $\langle\cdot\rangle_\beta$ is the expectation over $p_\beta(s' \mid c, \theta)$.

In practice, the method consists in first equilibrating the Gibbs sampler at $\beta = 0$, and then, performing a series of $K_{Th}+1$ cycles, where at each step, the value of β is increased by a small amount $\delta\beta = 1/K_{Th}$. The successive values of $\frac{\partial G}{\partial \beta}$ obtained during this quasi-static sampling scheme are recorded, and their average is our estimate of $\ln Y_1 - \ln Y_0$:

$$\ln Y_1 - \ln Y_0 \simeq \frac{1}{K_{Th}} \left[\frac{1}{2} \frac{\partial G(s_0, c)}{\partial \beta} + \sum_{h=1}^{K_{Th}-1} \frac{\partial G(s_h, c)}{\partial \beta} + \frac{1}{2} \frac{\partial G(s_{K_{Th}, c})}{\partial \beta} \right]. \quad (4.46)$$

Note that these developments are for one protein, but the generalization over the database is straightforward.

In the conditions of the present work, $K_{Th} = 1,000$ is sufficient to obtain an estimate of $\ln Y_1 - \ln Y_0$ with an error less than one natural unit of logarithm.

4.2.6.4 Model comparison

We measured the fit of each model using cross-validation (CV) : the potentials optimized on a first data set, i.e. the learning set, (θ_L) are applied on the second data set (the test set), and the log-likelihood is directly taken as a measure of fit. More precisely, for each model M,

$$CV_M = -\ln p(S_T \mid C_T, \theta_L, M), \quad (4.47)$$

where S_T and C_T are the sequences and structures of the test set. The difference with the CV score obtained for the flat potential (μ) is reported : $\Delta CV = CV_\mu - CV_M$.

4.2.6.5 Sequence sampling : site-specific profiles

Once an optimal value of θ is obtained, sequences compatible with a given conformation can be sampled from $p(s \mid c, \hat{\theta})$ by Gibbs sampling, and then further investigated. For instance, the frequency of each of the 20 amino acids (a) at each position (i) can be computed $(q_i(a))$, yielding a vector of site-specific *marginal* profiles, graphically displayed as sequence logos [Schneider and Stephens, 1990]. Alternatively, *leave-one-out* profiles can be obtained by computing the probability of each of the 20 amino-acids at each site of the test sequence, given the potential and the native sequence at all other positions :

$$p(s_i = a \mid s_j, j \neq i, \theta). \quad (4.48)$$

We measured the amount of information displayed by the profiles using the site-specific Shannon entropy :

$$h_i = -\sum_a q_i(a) \ln q_i(a) \quad (4.49)$$

We compared both marginal and leave-one-out profiles to the *empirical* profiles, i.e. profiles displayed by natural sequences. We generated these empirical profiles from multiple sequence alignments obtained from the ConSurf-HSSP database [Glaser et al., 2005].

4.2.6.6 Sequence sampling : Design specificity

As a test for specificity, designed sequences were submitted to a fold recognition experiment, using the fold recognition program THREADER [Jones et al., 1992c]. In THREADER, the compatibility of a sequence s for a given structure c is measured by the Z-score :

$$Z = \frac{\langle E(s,C)\rangle - E(s,c)}{\sigma} \tag{4.50}$$

where $\langle E(S,C)\rangle$ is the average of the THREADER statistical potential over all conformations of the decoy set, and σ is the corresponding standard deviation.

We randomly chose 70 structures of sizes ranging from 100 to 300 residues from the default THREADER dataset [see Additional file 8]. Structures whose native sequences produced a Z-score < 3 were discarded for the analysis. For each structure, c, we sampled 20 sequences from $p(s \mid c, \hat{\theta})$ by Gibbs sampling. These designed sequences were then submitted to THREADER [Jones et al., 1992c], and their specificity for the target structure c was measured by the ranking of c among all other structures, sorted by increasing Z-score.

A subset of 120 among the 1,200 sequences generated with the combined $(\epsilon+\alpha_{14ac}+\mu)$ potential (3-5 sequences for 23 distinct conformations, chosen at random ; [see Additional file 8]) were also submitted to another fold recognition program, LOOPP [Meller and Elber, 2001], and the presence of the native conformation c as the first hit or in the first 10 hits was recorded.

4.2.6.7 Learning databases

We used proteins culled from the entire PDB according to structure quality (resolution better than 2.0 Å) and with less than 25% of mutual sequence identity [Wang and Dunbrack, 2003]. Two subsets of approximately equal size were obtained by partitioning the proteins randomly : DS1, 449 proteins, 100,077 sites, and DS2, 465 proteins, 99,894 sites. The final list of proteins is available as supporting material[see Additional file 2][see Additional file 3].

4.2.7 Authors' contributions

CLK participated in the implementation of the methods, performed the run of all the experiments, and co-wrote the manuscript. NR participated in the implementation of the methods, extensively supervised CLK and CB, and made contributions to the drafting of the manuscript. CB participated in the initial implementation of the methods. HP contributed to the drafting of the manuscript and the coordination of the project. NL set up the theoretical framework, co-wrote the manuscript, participated in the implementation of the methods and directed the overall project. All authors read and approved the final manuscript.

4.2.8 Acknowledgements

Authors are grateful to Thomas Simonson, Pierre Tufféry, Laurent Chiche, Jérome Gracy and Gertraud Burger, for their critical comments on the manuscript and useful discussions. This work was financially supported in part by the "60ème comission franco-québécoise de coopération scientifique". CLK was supported by NSERC, CIHR and the Université de Montréal ; NR was supported by a bioinformatics grant from Génome Québec ; HP by the Canada Research Chair Program and the Université de Montréal ; CB and NL were funded by the french Centre National de la Recherche Scientifique, through the ACI-IMPBIO Model-Phylo funding program.

4.3 Conclusion

Cette première approche nous a permis de définir un cadre entièrement probabiliste pour l'optimisation de potentiels statistiques. Bien que la forme du potentiel présenté dans cet article semble simple au premier abord, l'approche proposée ici est intéressante. En effet, le potentiel optimisé dans cet article montrait déjà une amélioration très marquée par rapport au potentiel MJ [Miyazawa and Jernigan, 1985], bien que leur potentiel et le notre étaient censés incorporer les mêmes types d'information. Le test de validation croisée, présenté à la figure 4.4 est une preuve que les approximations induites par les modèles et les contextes dans lesquels les potentiels sont optimisés conduisent à de grandes différences en terme d'adéquation.

Cependant, la forme du potentiel et les approximations utilisées sont simples, et il serait intéressant de complexifier un peu ce potentiel, pour le rendre un peu plus réaliste, tout en restant dans les limites imposées par le modèle d'évolution (c'est à dire que le potentiel doit pouvoir être calculé rapidement). D'un côté, on peut vouloir intégrer des

termes supplémentaires, comme par exemple intégrer un potentiel dépendant des angles de torsion. D'un autre côté, on peut vouloir approximer le facteur de normalisation Z_s par une fonction plus réaliste que le *random energy model*.

Néanmoins, l'optimisation des potentiels faisait appel à une procédure lourde : à chaque étape du gradient, des séquences devaient être générées par MCMC, suivant la distribution de probabilité déterminée par les valeurs actuelles du potentiel (cf. Methods : Monte Carlo implementation). Cette étape était limitante pour la création et le test de nouvelles formes de potentiels. C'est pour cette raison que nous avons développé une méthode d'optimisation légèrement différente, basée sur une pseudo-vraisemblance, et qui est le sujet du chapitre suivant.

Chapitre 5

Optimisation des potentiels à l'aide d'une pseudo-vraisemblance

5.1 Introduction

Une fois le cadre statistique mis en place [Kleinman et al., 2006], les potentiels obtenus furent testés dans le cadre du modèle d'évolution soumis à des contraintes structurales [Rodrigue et al., 2009]. Cependant, il est vite apparu que ces potentiels ont une forme qui parait trop simple, et qu'il serait intéressant de tester de nouvelles formes de potentiel, tout en restant dans les contraintes imposées par le modèle. Malheureusement, la génération de séquences par MCMC pour le calcul du gradient était une étape limitante dans l'optimisation des paramètres. Ainsi, il fallait environ quinze jours pour estimer les paramètres (énergies) d'un potentiel statistique. Dans l'idée de construire des potentiels plus complexes, de nombreuses combinaisons possibles doivent être testées, afin ajuster progressivement les paramètres pour que le potentiel soit le meilleur possible. Il était donc nécessaire de réduire le temps de calcul nécessaire pour l'optimisation de chaque potentiel.

Tout le problème provient du facteur de normalisation de la vraisemblance, Y, défini dans l'équation suivante :

$$p(s \mid c, \theta) = \frac{e^{-G(s,c)}}{\sum_{s' \in \mathbb{S}} e^{-G(s',c)}} = \frac{e^{-G(s,c)}}{Y} \tag{5.1}$$

où $G(s, c)$ est le potentiel inverse et $\mathbb{S}$ représente l'ensemble des séquences. Le gradient du terme de normalisation est défini par :

$$\frac{\partial \ln Y}{\partial \theta} = -\sum_{s \in \mathbb{S}} \frac{\partial G(s,c)}{\partial \theta} p(s|c,\theta) = \langle \frac{\partial G(s,c)}{\partial \theta} \rangle. \tag{5.2}$$

Il s'agit donc d'une espérance sur $s \sim p(s|c,\theta)$ que l'on estimait par MCMC, en générant des séquences suivant la distribution de probabilité $p(s|c,\theta)$.

L'approximation inspirée du *leave-one-out* que nous présentons ici permet de contourner ce facteur de normalisation. Cette approximation avait déjà été évoquée dans [Kuhlman and Baker, 2000], sans qu'elle ait cependant été validée en tant qu'approximation alternative. Pour une raison, notamment : cette approximation suppose une dépendance à la séquence native qui pouvait fausser les résultats par rapport aux méthodes MCMC développées dans le chapitre précédent, qui sont indépendantes de cette séquence native.

Cet article nous permet de poser que ce problème potentiel ne se pose pas en pratique, et que l'approximation inspirée du *leave-one-out* permet d'obtenir des potentiels statistiques équivalents à ceux précédemment optimisés, mais avec un gain en temps de calcul remarquable (d'un facteur 1000) pour obtenir des potentiels avec des précisions équivalentes.

5.2 Fast optimization of statistical potentials for structurally constrained phylogenetic model

Authors : Cécile Bonnard, Claudia L. Kleinman, Nicolas Rodrigue, and Nicolas Lartillot.

5.2.1 Abstract

Background : Statistical approaches for *protein design* are relevant in the field of molecular evolutionary studies. In recent years, new, so-called structurally constrained (SC) models of protein-coding sequence evolution have been proposed, which use statistical potentials to assess sequence-structure compatibility. In a previous work, we defined a statistical framework for optimizing knowledge-based potentials especially suited to SC models. Our method used the maximum likelihood principle and provided what we call the *joint* potentials. However, the method required numerical estimations by the use of computationally heavy *Markov Chain Monte Carlo* sampling algorithms.

Results : Here, we develop an alternative optimization procedure, based on a *leave-one-out* argument coupled to fast gradient descent algorithms. We assess that the leave-one-out potential yields very similar results to the joint approach developed previously, both in terms of the resulting potential parameters, and by Bayes factor evaluation in a phylogenetic context. On the other hand, the leave-one-out approach results in a considerable computational benefit (up to a 1,000 fold decrease in computational time for the optimization procedure).

Conclusions : Due to its computational speed, the optimization method we propose offers an attractive alternative for the design and empirical evaluation of alternative forms of potentials, using large data sets and high-dimensional parameterizations.

5.2.2 Background

Recent advances in computer science and in the acquisition of new genetic sequences from a variety of organisms have opened up a wide spectrum of new possibilities in molecular evolutionary modeling. In particular, codon substitution models explicitly formulated in terms of a balance between mutation and selection constitute an attractive strategy [Halpern and Bruno, 1998, Yang and Nielsen, 2008, Rodrigue et al., 2009, Robinson et al., 2003]. By deriving the substitution process from basic principles of population genetics, their aim is to bridge the gap between population genetics and phylogenetics, and thus to offer a better understanding of the driving forces of the long term evolutionary process. More specifically, these mutation-selection models propose that the substitution rate from a sequence s to another s' ($R_{ss'}$) depends on the rate of mutation from s to s' ($Q^{mut}_{ss'}$), and on the probability for this mutation to be fixed in the population ($p_{fix}(ss')$) :

$$R_{ss'} = Q^{mut}_{ss'} \cdot p_{fix}(ss'). \tag{5.3}$$

The mutation matrix $Q^{mut}_{ss'}$ depends only on the underlying mutation model, and is generally assumed to be fixed along the lineages and uniform along the sequence. The fixation probability $p_{fix}(ss')$ depends on the particular model chosen.

Among the mutation-selection codon models, we focus on the structurally constrained (SC) models [Robinson et al., 2003, Rodrigue et al., 2005, Choi et al., 2007, Parisi and Echave, 2001] which attempt to explicitly link a protein's tertiary structure to the evolution of its sequence. They consider that a protein is under a purifying selection maintaining a stable and constant tertiary structure. Importantly, and unlike most probabilistic models currently used in molecular evolutionary studies, SC models are explicitly site-interdependent, and therefore, require complex Monte Carlo methods to be implemented and applied to empirical data [Robinson et al., 2003, Rodrigue et al., 2009, Choi et al., 2008].

In SC models, the fixation probability of a given mutation depends on a score function assessing the adequacy of a sequence s to the tertiary structure of the protein, c. This score should be devised so that the fixation probability is low if the proposed mutation destabilizes the structure or complicates the folding process. Since Anfinsen's experiments [Anfinsen, 1973], the relations between protein structure and sequence have

been carefully studied and an intuitive approach consists in relying on first principles of protein thermodynamics, using all-atom force fields (e.g. AMBER [Case et al., 2008], CHARMM [MacKerel Jr et al., 1998]). However, in our case, the instantaneous rate of substitution ($R_{ss'}$), and thus the structure/sequence score function, have to be computed for each possible nearest neighbor mutant, and for each substitution, along the entire evolutionary tree. Therefore, we need a fast computation of the fixation probability which precludes the use of all-atom force fields.

An attractive alternative is provided by knowledge-based (or statistical) potentials. They mimic the Boltzmann law [Miyazawa and Jernigan, 1985, Miyazawa and Jernigan, 1996, Sippl, 1993a, Solis and Rackovsky, 2006] and usually rely on a coarse-grained description of the structure, implicitly integrating out the degrees of freedom of the side chains and thus avoiding the complexity and the computation requirements of all-atom force fields [Tozzini, 2005, Seno et al., 1996, Deutsch and Kurowski, 1996, Seno et al., 1998, Rossi et al., 2000, Rossi et al., 2001, Moult, 1997, Mendes et al., 2002]. In addition, they are trained empirically from databases of natural proteins. This latter point is of particular interest in evolutionary studies, where we are interested in all aspects of the relations between sequence and structure prevailing in natural sequences, and not only in the specific problem of the thermodynamic stability. In this respect, one expects that learning potentials from native structure-sequence databases using blind machine learning methods will capture all such aspects.

Many statistical potentials have been proposed [Miyazawa and Jernigan, 1985, Bowie et al., 1991, Chiu and Goldstein, 1998b, Seno et al., 1998, Sippl, 1993a, Solis and Rackovsky, 2006], either to predict the fold of a given sequence (*protein folding*) or to find a sequence or a set of sequences folding into a given tertiary structure (*protein design*). However, the same potential may not be best-suited to both goals since the spaces of optimization are very different : in the protein folding problem the search is done over the structure space, while in the protein design problem the search is done over the sequence space. The phylogenetic context described here is more akin to a protein design perspective, as the structure of the protein is assumed constant during evolution, representing a constraint under which the sequence is evolving.

Several methods have been developed to train statistical potentials in a protein design perspective [Bowie et al., 1991, Seno et al., 1998, Chiu and Goldstein, 1998a]. In a previous work, we introduced a probabilistic framework for protein design purposes based on the maximum likelihood principle [Kleinman et al., 2006]. The likelihood we considered was the probability of the sequences S given their native structures C and the model parameters (here, the statistical potential parameters, θ), $P(S|C,\theta)$. This probability was then

maximized with respect to the potential parameters (e.g. pairwise contact energy coefficients) by a gradient method. However, the probability $P(S|C,\theta)$ involves a normalizing factor, summing over all possible sequences, which cannot be analytically calculated. We thus had to resort to a Markov Chain Monte Carlo ($MCMC$) numerical procedure : at each step of the gradient descent, we generated a set of sequences by Gibbs sampling, conditional on the current values of the potential. This set of sequences was then used to estimate the gradient. The Gibbs sampling procedure was the limiting step of our algorithm, restricting the set of alternative potentials that we could explore and empirically test. The potentials we obtained using this method are called *joint* potentials hereafter.

Interestingly, Kuhlman and Baker [Kuhlman and Baker, 2000] used a *leave-one-out* procedure to estimate a restricted set of parameters of a free physical energy function in order to do protein design. In this procedure, only one site of the protein is changed at a time, while the other positions are kept fixed in their native state. The procedure is thus similar to training a potential to recognize acceptable sequence variants, given the target structure, among all possible point mutants. The leave-one-out criterion seems to give good results. However, it has never been assessed against alternative methods. Here, we adapt the statistical framework we defined in [Kleinman et al., 2006] now using the leave-one-out definition of the likelihood to perform the gradient descent instead of the joint likelihood. We compare the potential parameters obtained by the two methods, and we establish that we can be highly confident in the results obtained using the leave-one-out likelihood. Overall, the leave-one-out procedure allows much faster computations while giving sensibly the same results as the joint one.

5.2.3 Results

5.2.3.1 Likelihood framework

As in [Kleinman et al., 2006], we formulate the problem in terms of a probabilistic model, considering a sequence $s = (s_i)_{1..n}$ of length n according to a probability distribution $P(s|c,\theta)$, conditional on the conformation c and on a set of potential parameters θ. The parameters are estimated by maximizing the probability of observing a database of N independent sequence-structure pairs $(\tilde{S}, C)$, with $\tilde{S} = (\tilde{s}^p)_{p=1..N}$, $C = (c^p)_{p=1..N}$. Here, $\tilde{s}^p = (\tilde{s}_i)_{i=1..n_p}^p$ is the p-th native sequence of the dataset, n_p is the lenght of this sequence and c^p is the native conformation associated with $\tilde{s}^p$. In practice, a native sequence-structure pair corresponds to a protein taken from the PDB.

The probability that we want to maximize can be expressed as follows :

$$P(\tilde{S}|C,\theta) = \prod_p P(\tilde{s}^p|c^p,\theta). \tag{5.4}$$

As a function of θ, this term can be seen as a likelihood. Hereafter, we define the methodology with one protein, but it can be easily generalized to a set of proteins.

Borrowing from [Kleinman et al., 2006], we set :

$$P(s|c,\theta) = \frac{e^{-G(s|c,\theta)}}{\sum_{s'\in\mathbb{S}} e^{-G(s'|c,\theta)}} = \frac{e^{-G(s|c,\theta)}}{Y}, \tag{5.5}$$

where Y is called the *normalization factor*, and $G(s|c,\theta)$ the *inverse potential*, defined as

$$G(s|c,\theta) = E(s|c,\theta) - F(s), \tag{5.6}$$

where $E(s|c,\theta)$ is the statistical potential and $F(s)$ is analogous to a free energy term and can be approximated using the *random energy model* [Shakhnovich and Gutin, 1993, Sun et al., 1995, Seno et al., 1998, Pande et al., 1997] :

$$F(s) = \sum_{1\leq i\leq n} \mu_{s_i}, \tag{5.7}$$

where $\mu_a, a = \{1..20\}$ are unknown parameters, analogous to *chemical potentials* [Kleinman et al., 2006].

5.2.3.2 Optimization method

Joint likelihood maximization

In our previous work [Kleinman et al., 2006], we defined a score function $\omega(\tilde{s}|c,\theta)$ as :

$$\omega(\tilde{s}|c,\theta) = -\ln P(\tilde{s}|c,\theta) = G(\tilde{s}|c,\theta) + \ln Y. \tag{5.8}$$

This score function should be minimized conditional to θ. Its gradient is :

$$\frac{\partial\omega(\tilde{s}|c,\theta)}{\partial\theta} = \frac{\partial G(\tilde{s}|c,\theta)}{\partial\theta} + \frac{\partial \ln Y}{\partial\theta} = \frac{\partial G(\tilde{s}|c,\theta)}{\partial\theta} - \langle\frac{\partial G}{\partial\theta}\rangle, \tag{5.9}$$

where $\langle\cdot\rangle$ stands for the expectation over sequences drawn from the probability defined by eq. 5.5. Given the size of the sequence space (20^n), this expectation cannot be computed analytically, and therefore, in [Kleinman et al., 2006] we used a MCMC method to numerically estimate this expectation.

Leave-one-out likelihood maximization

We define for site i, $i = 1..n$, the leave-one-out probability

$$P_i(a|\tilde{s}_{\setminus i}, c, \theta) = P_i(a|\forall\ j \neq i\ s_j = \tilde{s}_j, c, \theta), \tag{5.10}$$

which is the probability of having an amino acid a at site i, in the context of the native sequence at all other sites ($\forall j \neq i\ s_j = \tilde{s}_j$). This leave-one-out probability can easily be obtained by a normalization over all possible twenty outcomes at site i :

$$P_i(a|\tilde{s}_{\setminus i}, c, \theta) = \frac{e^{-G_i(s_i=a|\tilde{s}_{\setminus i}, c, \theta)}}{\sum_{k=1}^{20} e^{-G_i(s_i=k|\tilde{s}_{\setminus i}, c, \theta)}}. \tag{5.11}$$

We can write this probability for any amino acid a, and in particular for the native amino acid at site i, $\tilde{s}_i$ i.e. $p_i(\tilde{s}_i|\tilde{s}_{\setminus i}, c, \theta)$. Taking the product over all positions $i = 1..n$, and by analogy with our previous definition of likelihood, we define the leave-one-out likelihood :

$$P^l(\tilde{s}|\tilde{s}, c, \theta) = \prod_{1 \leq i \leq n} P_i(\tilde{s}_i|\tilde{s}_{\setminus i}, c, \theta). \tag{5.12}$$

Note that this leave-one-out likelihood is normalized over the sequences, exactly as in the case of eq. 5.5. Therefore it yields a valid probability distribution over the sequence space. On the other hand, the probability depends not only on c and θ, but also, in some sense, on the native sequence itself. To make this point explicit, we make $\tilde{s}$ appear on both sides of the conditioning bar.

We define the corresponding scoring function :

$$\omega^l(\tilde{s}|\tilde{s}, c, \theta) = -\ln P^l(\tilde{s}|\tilde{s}, c, \theta), \tag{5.13}$$

the gradient of which is immediately obtained (Additional File 1) :

$$\frac{\partial \omega^l(\tilde{s}|\tilde{s}, c, \theta)}{\partial \theta} = \sum_{i=1..n} \frac{\partial G_i(\tilde{s}_i|\tilde{s}_{\setminus i}, c, \theta)}{\partial \theta} - \sum_{i=1..n} \sum_{a=1..20} p_i(a) \frac{\partial G_i(a|\tilde{s}_{\setminus i}, c, \theta)}{\partial \theta}. \tag{5.14}$$

This gradient can be analytically calculated, at each step of a gradient descent. We note that the term corresponding to the normalization factor (the second term in eq. 5.14) can be seen as an expectation over the leave-one-out probability. It is thus analogous to the expectation appearing in the right hand of eq. 5.9. However, it is defined on a much more restricted universe ($20 \cdot n$ states, compared to the 20^n states in the case of the joint likelihood).

For implementing both methods, we used a simple form of potential [Kleinman et al., 2006], consisting in two terms : one related to contact interactions and the other to the solvent accessibility (see Methods).

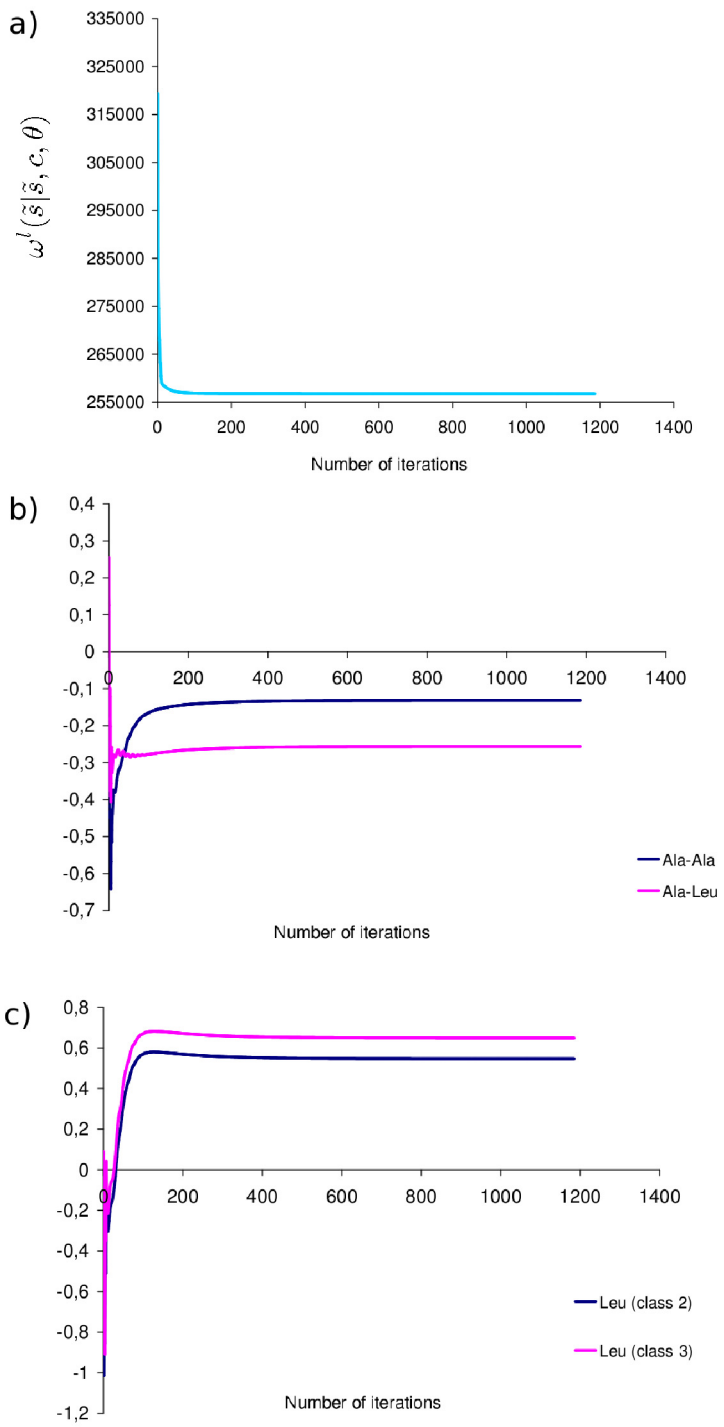


FIGURE 5.1 – Convergence of the optimization procedure. - Evolution of (a) the score function, (b) contact potential parameters and (c) accessibility potential parameters, for the dataset DS_l, using the controlled inertial gradient descent.

5.2.3.3 Potential optimization

We first run our leave-one-out method on DS_l (see Methods). We consider that the optimization is complete when the overall maximum gradient is smaller than 10^{-2}. This corresponds to a variation of 10^{-6}, at most, in the value of the potential parameters. Using this stopping condition on the dataset DS_l with empirically tuned general steps (e.g for the contact parameters : $\delta^c_{grad} = 10^{-5}$ and for the solvent accessibility parameters : $\delta^a_{grad} = 10^{-4}$), we compare three different gradient descent methods (described in Methods) : the simple gradient descent, the inertial gradient descent, and the controlled inertial gradient descent. The values of the parameters stabilized after 14,500 gradient steps for the simplest gradient descent, versus 1,500 gradient steps for the inertial gradient, and 1,200 gradient steps for the controlled inertial gradient. Concerning the last method, if we choose a different general step (e.g. $\delta^c_{grad} = 10^{-3}$ and $\delta^a_{grad} = 10^{-2}$) the procedure automatically reaches the optimal step for that dataset. At the beginning of the optimization procedure, the inertial component of the gradient greatly speeds up the optimization, but is automatically deactivated when the values of the potential parameters are near the optimum, thus avoiding the numerical instabilities usually observed using less adaptive gradient methods.

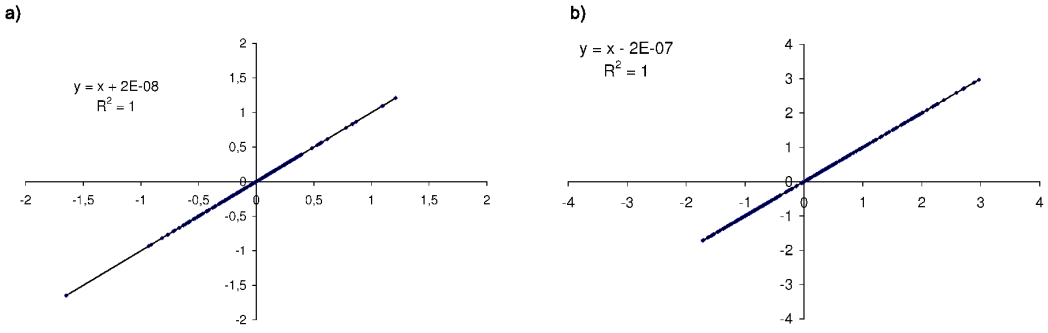


FIGURE 5.2 – XY comparisons of the leave-one-out potential parameters. - XY comparisons of two independent runs on the same dataset DS_l for (a) contact and (b) solvent accessibility potential parameters respectivly.

Independent runs from different and randomly chosen initial values for the parameters of the leave-one-out potential (θ^l), lead to the same final values of $\omega^l(\tilde{s}|\tilde{s}, c, \theta)$ (fig. 5.1) and of the potential parameters (fig. 5.2). These computations were done with the three gradient descent methods, and resulting always in the same final values, which suggests

that, in the present case, we do not have local minima in the space of parameters. Similarly, the potential parameters obtained by two independent runs on the same dataset are very similar, indicating that our stopping condition is sufficient to have a good precision in our estimates (Additional file 2). In fig. 5.1 we have also represented the evolution of some parameters of the potential during optimization. We can see that the values of these parameters oscillate at the beginning of the gradient descent and then reach their optimal values. This behavior is caused by the evolution of the other parameters, as they influence each other during optimization. The complete series of parameter values obtained by our optimization method are presented in the additional file 3.

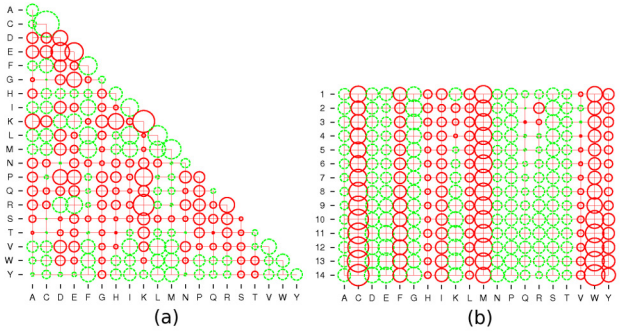


FIGURE 5.3 – Validation of the potential parameters. - Bubble plot representations of (a) contact potential parameters and (b) accessibility potential parameters obtained upon the dataset DS_l. Negative values are plotted in green while positive values are plotted in red.

The contact potentials obtained with the leave-one-out optimization criterion make sense from a biological point of view (fig. 5.3) : as expected, favorable interactions between amino acids in the contact potentials are represented by large negative value (e.g. the Cysteine-Cysteine contact energy, fig. 5.3), and by large positive value for unfavorable interactions (e.g. the Lysine-Lysine or Lysine-Arginine interactions, which are electrostatically repulsive). Concerning the accessibility potentials, it is important to note that we are working in a protein design context (i.e. we are evaluating the fitness of alternatives amino acids in a given accessibility class). Accordingly, the accessibility potentials have to be interpreted row-wise. If one wants to compare the accessibility potentials between classes for a given amino acid (i.e. in a protein folding perspective), one solution is to remove the logarithm of the frequency of the accessibility classes to each potential (additional file 4). Also, note that there is a lack of identifiability between α and μ, which has been be resolved by including the chemical potentials in the accessibility terms.

5.2.3.4 Complexity

In our previous work, we had to use a MCMC protocol to numerically evaluate the derivative of the gradient (see. eq. 5.9), which was a computationally demanding task. At each step of the gradient descent, we had to sample a set of sequences by Gibbs sampling, under the current values of the parameters, so as to numerically estimate the gradient of the log-likelihood.

To compare the joint and the leave-one-out potentials, we first define an elementary calculation as the evaluation of the *inverse* potential at a particular site i for one particular amino acid a (what we called $G_i(a|\tilde{s}_{\setminus i}, c, \theta)$, eq. 5.11). This calculation has to be made in both cases. It is explicitly defined in the leave-one-out procedure (eq. 5.12), and is implicitly used in the joint context : an elementary step of the Gibbs sampling algorithm consist in computing, at a given site i the leave-one-out probability (eq. 5.11) for each possible amino-acid at this site, conditional on the rest of the sequence, and to choose the new amino-acid at site i according to these probabilities. Performing such an elementary update for every site in turn corresponds to one Gibbs sampling sweep and represents $20 \cdot n$ elementary computations. A reliable estimate of the joint expectation requires K sweeps (burn in included) and so, for a gradient step, we need $K \cdot n \cdot 20$ elementary calculations (in practice, $K \simeq 1,000$).

In the case of the leave-one-out potential, we only have to make the equivalent of one sweep to exactly compute the gradient (eq. 5.14). Thus, we only need $n \cdot 20$ elementary calculations for a gradient step, which thus represents a 1,000-fold increase in computational speed compared to the joint method. In practice, and after the addition of the acceleration of the gradient descent, it took about one week to have a good estimate when we used the joint criterion, versus less than fifteen minutes when using the leave-one-out approach.

5.2.3.5 Potentials are indistinguishable

We applied the two optimization procedures (joint and leave-one-out) to the same dataset DS_j, and found a high correlation between the two resulting potentials (fig. 5.4). The correlation coefficient R^2 was about 0.96779 for the contact potential parameters and about 0.97374 for the accessibility potential parameters. For comparison, we applied the leave-one-out procedure on the two datasets DS1 and DS2 (see additional file 2) and found a correlation coefficient of 0.9477 for the contact parameters and of 0.9596 for the accessibility parameters, indicating that the difference between the joint and the leave-one-out potentials is small compared to the sampling error due to the finite size of the training

set. Altogether, the leave-one-out method appears to be a fast and reliable optimization procedure, yielding potentials that are virtually indistinguishable from those obtained under the joint method. As presented in [Kleinman et al., 2006], the contact potentials present a correlation ($R^2 = 0.6565$) with those of Miyazawa and Jernigan [Miyazawa and Jernigan, 1996].

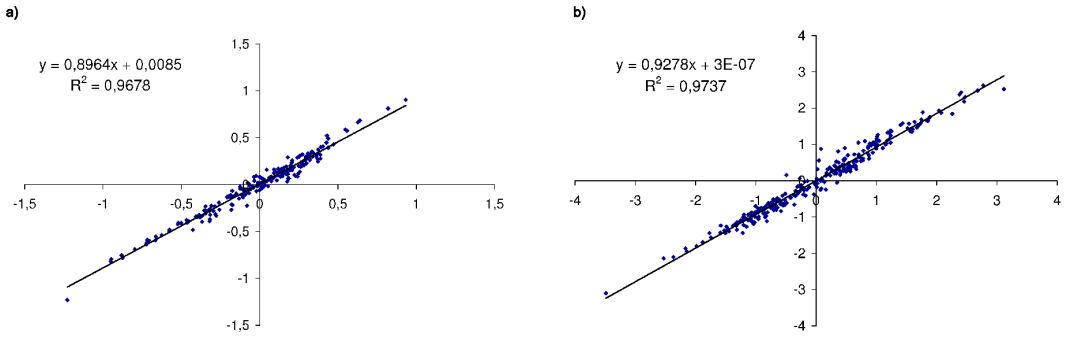


FIGURE 5.4 – XY comparisons of the leave-one-out and joint potential parameters. - XY comparisons between the two potentials (optimized on the same dataset DS_j), with, in X-axis the leave-one-out potential, and in Y-axis the joint potential. (a) represents the correlation between the contact potential parameters, and (b) the correlation between the accessibility potential parameters.

5.2.3.6 Phylogenetic evaluation

In eq. 7.4, we defined the substitution process of the SC model as a process depending on a mutation rate and a fixation probability. There are many ways the fixation probability could be expressed. Here, we do as in Robinson et al [Robinson et al., 2003] and assume that this probability depends only on the potential difference (ΔG) between the original and the mutated sequences. Let us denote by s_{nuc} and s'_{nuc}, two sequences which differ only by a nucleotide, and s_{aa} and s'_{aa}, the corresponding amino acid sequences (which may be identical due to codon synonymy). Then, the rate of substitution between s and s' is :

$$R_{s_{nuc}s'_{nuc}} = Q^{mut}_{s_{nuc}s'_{nuc}} \cdot e^{-\beta \Delta G_{s_{aa}s'_{aa}}}, \tag{5.15}$$

where $Q^{mut}_{s_{nuc}s'_{nuc}}$ is the mutation term depending only on the two sequences s_{nuc} and s'_{nuc}. $\Delta G_{s_{aa}s'_{aa}}$ is the energy difference between s_{aa} and s'_{aa}, and $\beta \geq 0$ can be considered as

the strength of the structure-sequence constraint enforced by the model. Thus, a negative (resp. positive) ΔG means that the mutation is more (resp. less) likely to be accepted than a purely neutral (e.g. synonymous) mutation.

Note that the substitution process defined by eq. 5.15 is reversible and has a stationary distribution defined by :

$$\Pi_s \propto \Pi_0(s_{nuc}) e^{-2\beta G(s_{aa})}, \tag{5.16}$$

where $\Pi_0(s_{nuc})$ is the stationary distribution induced by the pure mutation process ($Q^{mut}_{s_{nuc}s'_{nuc}}$). Given the way our potentials are optimized (see eq. 5.5 and 5.11) and assuming that natural sequences are sampled at equilibrium from the process defined by eq. 5.15, we then expect that the optimal value of β should be close to 0.5. In the following, we explore the entire range $\beta \in [0, 1]$.

We denote by SC^l_β the SC model defined using the leave-one-out potential and SC^j_β the SC model defined using the joint potential ; the two models depend on β. Obviously, when $\beta = 0$, $SC^l_0 = SC^j_0 = SC_0$, and the model reduces to a pure mutation model which will be considered as our reference model.

We implemented our potential in the SC model as described in [Rodrigue et al., 2009] and applied it to the GLOBIN15-144 dataset, with an underlying mutational specification inspired by the codon model in [Muse and Gaut, 1994] and denoted as MG in [Rodrigue et al., 2009]. This MCMC framework allows one to obtain a sample of parameter values and substitutional histories along the tree, drawn from the posterior distribution under the $SC^l_{0.5}$ model. Such a sample can then be marginalized over quantities of interest. Here, we briefly illustrate the approach by displaying the logo of the reconstructed mammalian ancestor hemoglobin sequence (fig. 5.5).

Since the leave-one-out procedure can be seen as an approximate but faster training method, compared to the joint method developed previously, we evaluated its impact on model fit via Bayes factors evaluations (see Methods). In this section we consider the three versions of the SC model, SC^l_β, based on a contact + accessibility leave-one-out potential, SC^j_β, based on a contact + accessibility joint potential, and SC^c_β based on a contact only joint potential. As explained in the methods, in the present case, the thermodynamic integration method yields a complete fitness curve (fig. 5.6) of each model (i.e. a curve representing the Bayes factor of each model against the reference model, as a function of β). In this way, we can readily spot the optimal value of β under each model, and report the Bayes factors under this optimal value (table 5.1).

As can be seen from fig. 5.6 and table 5.1, the models based on the joint and the leave-one-out potentials have a very similar fit across the whole range of value of β that we tested. Interestingly, in all but one cases, the Bayes factor appears to be slightly in favor

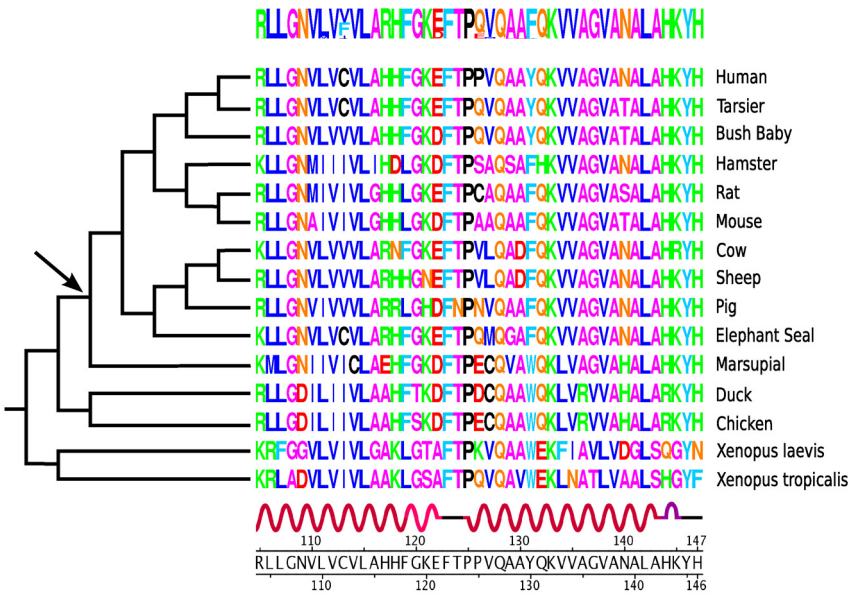


FIGURE 5.5 – Logo profile of the mammalian ancestral globin sequence. - The node is marked by an arrow. The translated sequences of the true alignment are displayed along with the secondary structure of the structure PDB code 4HHBB.

of the leave-one-out potential, although the differences are not significant. As a point of comparison, we also measured the fit of the contact only potential (joint method), to illustrate that the difference between the joint and the leave-one-out methods is small compared to the differences observed between the alternative forms of statistical potential that we would like to empirically compare (see [Kleinman et al., 2006] for an evaluation of the relative contribution of each potential component to the fitness of the model).

	ADH23-254	CALM36-444	GLOBIN15-144	LYS25-134
SC_β^c	[74.748-75.032]	[149.819-149.929]	[57.953-58.135]	[11.5-11.968]
SC_β^j	[102.666-102.766]	[161.340-161.491]	[70.666-70.948]	[26.287-26.417]
SC_β^l	[102.977-103.115]	[158.679-158.858]	[72.485-72.872]	[29.545-29.852]
optimal β	[0.387-0.397]	[0.371-0.383]	[0.450-0.498]	[0.179-0.249]

TABLE 5.1 – The natural logarithm of the Bayes factors.

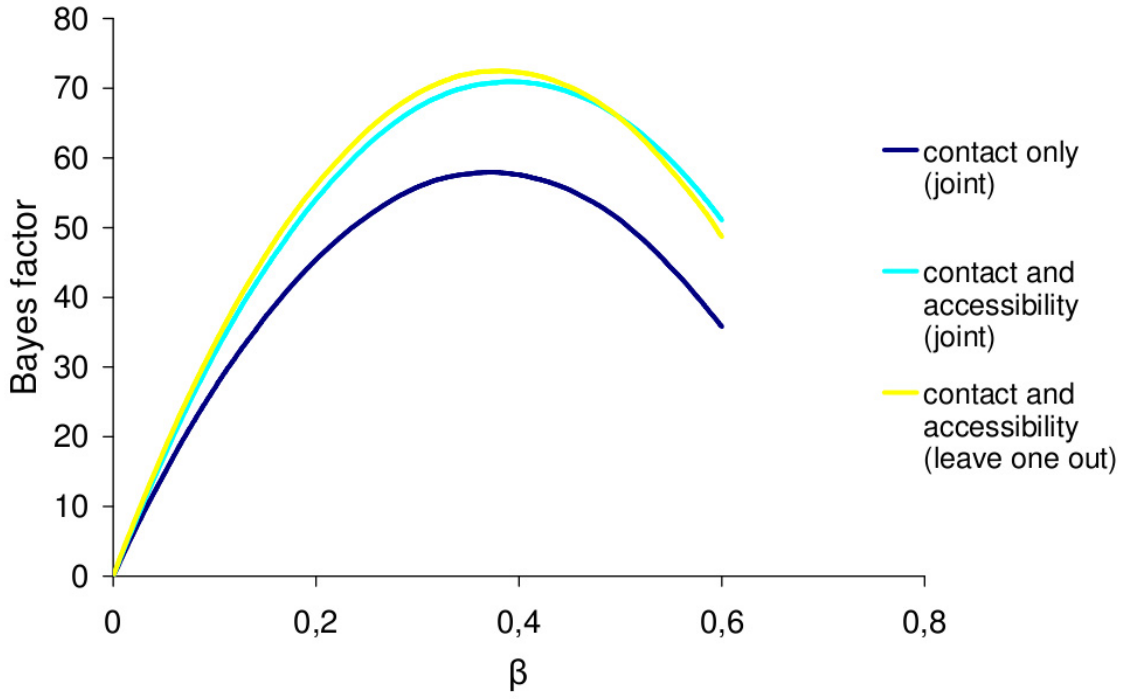


FIGURE 5.6 – Bayes factor. - Curves representing the Bayes factor as a function of β, with SC^l_β (in yellow), SC^j_β (in light blue) and SC^c_β (in dark blue), for the dataset BGLOBIN15-144.

5.2.4 Discussion

In a previous work [Kleinman et al., 2006], we defined a statistical framework for protein design, using the maximum likelihood principle, with the aim of devising statistical potentials to be used in phylogenetic studies. However, the optimization procedure we introduced at that time requires a MCMC protocol to cope with the proportionality constant entailed by the normalization of the probability over the sequence space. Here, we introduce a different likelihood, which we called leave-one-out, to optimize the potentials. A similar procedure was previously used by Kuhlman and Baker [Kuhlman and Baker, 2000], but was not statistically assessed against alternative procedures. We found in this work that the joint and the leave-one-out potentials are virtually indistinguishable, both by direct comparison and by Bayes factor evaluation in a phylogenetic context.

We note that the optimal β for the SC^l_β model is not 0.5, as one may expect given the way our potentials were normalized (see eq. 5.5, 5.8 and 5.15). Several explanations can be proposed. First, strictly speaking, this expectation is valid under the joint procedure, and not under the leave-one-out procedure. But the very high similarity between the two

resulting potentials, and the fact that a similar phenomenon ($\beta \neq 0.5$) can be observed also under a potential optimized using the joint method [Rodrigue et al., 2009] do not favor this explanation. Alternatively, it may appear at first that this could be due to the fact that the underlying mutation model (the Q^{mut} matrix in eq. 5.15) was not explicitly taken into account when optimizing the potential (so that the chemical potentials implicitely include a mutational component), whereas our phylogenetic model does involve an explicit mutational process. In this sense, in the phylogenetic analysis, there is a potentially (partially) redundant modeling of mutational features, in having explicit parameters devoted to these, in combination with the use of the SC setting. This might explain the optimal value of β lower than 0.5. The phenomenon may also be the result of model violations, which are very likely to be present given the simple form of the potentials. Finally, it is also likely that the mutation pressure, or the selection strength (represented by β) is not the same for each protein. Accordingly, two possible improvements to the method can thus be proposed here : the first is to optimize the potential while allowing for different values of β for each protein or each family of protein. The second is to cluster proteins into classes, and optimize a potential specifically for each class.

5.2.5 Conclusions

Apart from these two possible improvements, many other directions of research should now be explored : alternative functional forms for the potential should be implemented and empirically tested. Several methods accounting for negative design, through the use of explicit decoys [Deutsch and Kurowski, 1996] such as the use of a normalized energy gap between a native structure and misfolded structures [Bastolla et al., 2006], or using variational methods [Seno et al., 1998], also deserve further investigation. The supervised learning described here depends on structure-sequence pairs. In the present case, we have used native pairs, but this could be relaxed by taking a set of structures (e.g. obtained by molecular dynamics) as the reference structure or by taking a set of homologous sequences instead of a unique sequence [Panjkovich et al., 2008]. A more appealing method would consist in doing the optimization directly within the phylogenetic context. Importantly, the fact that the leave-one-out procedure is much faster than the joint method (in the present case, roughly by a factor $1,000$), has obvious practical consequences, as it allows a much larger diversity of alternative models and methods to be tested.

5.2.6 Methods

5.2.6.1 Gradient descent

When performing a gradient descent, several methods can be used. We expose here the three gradient descent methods that we compared. In all cases, the method rely on a cyclical updating of parameter values, where, given the values of parameters at the m^{th} cycle, which we write as $\theta^{(m)}$, the update is given by :

$$\theta^{(m+1)} = \theta^{(m)} - \Delta\theta^{(m+1)}. \tag{5.17}$$

The increment, $\Delta\theta^{(m+1)}$, is conditional to the scoring function, that we simply denote in this part as $\omega(\theta^{(m)})$.

Fixed step gradient

This is the simplest form of the gradient descent. We write :

$$\Delta\theta^{(m+1)} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m)})}{\partial\theta}, \tag{5.18}$$

where δ_{grad} is the fixed step of the gradient descent. Even though this formalism is simple, the choice of the step is not trivial. Indeed, if the step is too large, the values of the potential will oscillate around the optimal values. Conversely, if the step is too small, the gradient descent will be too slow.

Inertial gradient

To reduce the optimization time, another method of gradient descent was developed, based on an analogy with the physical phenomenon of inertia.

$$\Delta\theta^{(m+1)} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m)})}{\partial\theta} + \delta_{iner} \cdot \Delta\theta^{(m)}. \tag{5.19}$$

δ_{iner} is the damping rate of the inertial component, $0 \leq \delta_{iner} < 1$. If $\delta_{iner} = 0$, eq. 5.19 reduces to the case of the simple gradient. In practice, we set δ_{iner} equal to 0.9.

However, there is a drawback when taking into account the previous variation of the parameters : when the directions of the gradient change, the inertial part of the gradient brings the parameters too far beyond the maximum. In addition, the gradient step δ_{grad} has to be small enough so that the values of the potential do not oscillate around the optimal values, as in the case of the fixed step gradient.

Controlled inertial gradient

To avoid these two drawbacks, we define here a controlled inertial gradient descent formalism. Specifically, let us define :

$$\Delta\theta^{\star} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m-1)})}{\partial\theta} + \delta_{iner} \cdot \Delta\theta^{(m)}, \tag{5.20}$$

$$\Delta\theta^{\bullet} = \delta_{grad} \cdot \frac{\partial\omega(\theta^{(m-1)})}{\partial\theta}. \tag{5.21}$$

The decision procedure can thus be described as follows (see additional file 5). First, we test if the addition of $\Delta\theta^{\star}$ (derivative component and inertial component) to the actual values of parameters $\theta^{(m)}$ gives a higher likelihood than $\theta^{(m)}$. If it does, then the step corresponds to a classical step of the inertial gradient descent. Otherwise, the algorithm tests if the addition to $\theta^{(m)}$ of the derivative component ($\Delta\theta^{\bullet}$) only gives a higher likelihood than the actual values. If it does, the step corresponds to a classical gradient descent. Otherwise, we retry a simple gradient descent with a smaller δ_{grad}.

The above procedure has two advantages. The first is the speed-up offered by the inertial component, when its addition has a positive influence on the likelihood. The second advantage is that the last part of the algorithm automates the search for an optimal value of the steps, and avoids both oscillations of θ around the optimum, and a slow gradient descent.

5.2.6.2 Statistical potentials

We used the same statistical potential function as in our previous work [Kleinman et al., 2006]. The (pseudo) energy score consists of two terms :

$$E(s|c) = \sum_{1\leq i<j\leq n} \Delta_{ij}\varepsilon_{s_i s_j} + \sum_{1\leq i\leq n} \alpha_{s_i}^{\nu_i}. \tag{5.22}$$

The first term represents the contact free energy (defined between sidechain centers) : $\Delta_{ij} = 1$ if i and j are closer than the cutoff distance (here 6.5 Å), and ε_{ab} represents the contact potential between amino acids a and b. The second term represents the accessibility free energy : ν_i is the accessibility class of the site i and α_a^d is the solvent accessibility potential of the amino acid a when placed into the accessibility class d ($d = \{1..D\}$), where D is the number of accessibility classes.

We use the *random energy model* principle to approximate $F(s)$ (eq. 5.7), so that the inverse potential becomes :

$$G(s|c,\theta) = \sum_{1\leq i<j\leq n} \Delta_{ij}\varepsilon_{s_i s_j} + \sum_{1\leq i\leq n} \alpha_{s_i}^{\nu_i} + \sum_{1\leq i\leq n} \mu_{s_i}. \tag{5.23}$$

As in our previous work we fix the constraints :

$$\sum_{1\leq a\leq 20} \mu_a = 0, \tag{5.24}$$

$$\sum_{1\leq a\leq 20} \sum_{1\leq b\leq 20} \varepsilon_{ab} = 0, \tag{5.25}$$

$$\sum_{1\leq a\leq 20} \alpha_a^d = 0, d = \{1..D\}, \tag{5.26}$$

since $G(s|c,\theta)$ is invariant under the following transformations $\mu'_a = \mu_a + J_1$, $\varepsilon'_{ab} = \varepsilon_{ab} + J_2$ and $\alpha_a'^d = \alpha_a^d + J_3$. However, there is an additional lack of identifiability between α and μ, which can be resolved by including the chemical potentials in the accessibility terms. Indeed, the μ_a terms can be seen as an additive constant to each accessibility term for a given accessibility class (see additional file 6). In the present case, our final inverse potential is therefore :

$$G(s|c) = \sum_{1\leq i<j\leq n} \Delta_{ij}\varepsilon_{s_i s_j} + \sum_{1\leq i\leq n} \alpha_{s_i}^{\nu_i}, \tag{5.27}$$

and our set of parameters for the statistical potential will thus consist of :

$$\theta = \{\varepsilon_{ab}, \alpha_a^d\},\ 1 \leq a \leq 20,\ 1 \leq b \leq 20,\ d = \{1..D\}. \tag{5.28}$$

5.2.6.3 Bayes factor evaluation

In a Bayesian statistical framework the method of choice for comparing models is to compute Bayes factors. The Bayes factor between two models is defined as the ratio of their respective marginal likelihood. The case $B(SC_0, SC_\beta^l) > 1$ (resp. $B(SC_0, SC_\beta^l) < 1$) is considered as an evidence in favor of (resp. against) the SC_β^l model. We write the Bayes factor between SC_0 and SC_β^l as :

$$B(SC_0, SC_\beta^l) = \frac{P(A|SC_\beta^l)}{P(A|SC_0)}, \tag{5.29}$$

where A corresponds to the data, composed by an alignment of coding nucleotide sequences and a topology and

$$P(A|SC_\beta^l) = \int_\theta P(A|\theta)P(\theta)d\theta. \tag{5.30}$$

Here we compute Bayes factors by thermodynamic integration (or *path sampling*) as described in [Rodrigue et al., 2009]. The procedure consists in sampling along a continuous

path between SC_0 and SC_β^l through a set of slight changes in the value of β. In fact, this procedure provides a complete curve representing the fit of the model as a function of β. Sampling from $\beta = 0$ to $\beta = \beta_{max}$ and from $\beta = \beta_{max}$ to $\beta = 0$ gives two different curves for the logarithm Bayes factor, which we used as an internal check of the reliability of the method (not shown).

5.2.6.4 Datasets

Optimization datasets

The datasets are made of proteins (structure-sequence pairs) culled from the PDB, with less than 25 % of mutual sequence identity and a resolution better than 2 Å [Wang and Dunbrack, 2003]. This sequence homology percentage and the size of the database avoid possible bias that could be induced by related proteins. To compare the joint and leave-one-out potentials, we used the dataset on which we previously estimated the joint potentials, DS_j. This dataset is made of 441 proteins and 98,155 sites [Kleinman et al., 2006]. We also consider a dataset DS_l (made of 3,363 proteins and 835,717 sites) which was split into two subsets : $DS1$ (1,691 proteins and 419,208 sites), and $DS2$ (1,672 proteins and 416,509 sites). To determine the accessibility classes, we first compute the solvent accessibility area using Naccess 2.1 [Hubbard and Thornton, 1993] and partitioned the resulting values into classes [Kleinman et al., 2006].

Phylogenetic Datasets

The SC model was applied to 4 distinct multiple sequence alignments : Globin15-144, Lysin25-134, Adh23-254 and Calm33-444. Globin15-144 is made of 15 vertebrates sequences of the β-globin gene (taken from the original dataset from [Yang et al., 2000a]), with a protein structure defined by the PDB file 4HHB and a tree topology estimated using Phylobayes 3.1c [Lartillot et al., 2009] (which is consistent with the tree topology described in [Murphy et al., 2001]). Lysin25-134 is made of 25 Abalone sperm lysin sequences [Yang et al., 2000b], with a protein structure defined by the PDB file 1LYS and the tree topology previously defined by [Yang et al., 2000b]. Adh23-254 is made of 23 alcohol dehydrogenase sequences taken form Drosophila [Yang et al., 2000a], with a protein structure defined by the PDB file 1A4U and the tree topology previously defined by [Yang et al., 2000a]. Calm36-444 is made of 36 calmodulin sequences taken from eukaryotes, with a protein structure defined by the PDB file 1CFD and the tree topology estimated using phyML [Guindon and Gascuel, 2003] under the model JTT + F + Γ [Jones et al., 1992b, Yang, 1993].

5.2.7 Authors contributions

CB implemented the leave-one-out and gradient descent methods described here and performed the run of all the experiments. CLK implemented the data pre-processing methods. NR implemented the phylogenetic framework. NL set up the theoretical framework and directed the overall project. All the authors co-wrote the manuscript and approved the final manuscript.

5.2.8 Acknowledgements

The authors are grateful to the three anonymous referees for their useful comments on the manuscript. CB was financially supported by the french Centre National de la Recherche Scientifique (CNRS), the Région Languedoc-Roussillon and the Université de Montréal, CLK by NSERC, CIHR and the Université de Montréal, NR by NSERC, and NL by the Université de Montréal, NSERC and the CNRS.

5.3 Conclusion

J'ai dessiné dans la figure 5.7 les diagrammes représentant les deux algorithmes d'optimisation présentés précédemment : la méthode exacte (permettant l'obtention du *joint potential*, à l'aide des MCMC) à gauche, et la méthode de pseudo-vraisemblance (donnant lieu à l'optimisation du *leave-one-out potential*) à droite. L'intégration d'une fonction de pseudo-vraisemblance permet de simplifier l'algorithme d'optimisation en supprimant une boucle, tout en permettant d'obtenir des valeurs de potentiel semblables à celles obtenues précédemment.

Le cadre statistique et l'algorithme d'optimisation maintenant bien définis, il nous était alors possible de tester de nouvelles formes de potentiels, ou de complexifier l'approximation du Z_s par de meilleures approches que celle inspirée du *random energy model*, ainsi que nous en parlerons dans le chapitre 7.

De nouvelles formes de potentiels ont donc été déterminées, et testées, incluant des matrices de distances au lieu des matrices de contact, ainsi que des paramètres de torsion, de structure secondaire et même un terme essayant de prendre en compte la flexibilité au niveau des résidus [Kleinman et al., Submitted]. Chacun des nouveaux termes ainsi ajoutés au potentiel est construit sur le même modèle que le terme d'accessibilité au solvant : d'abord, un descripteur structural est défini (e.g. surface d'accessibilité au solvant, angle de torsion...) ; puis l'ensemble des valeurs possibles pour ce descripteur est partitionné en classes ; enfin, un vecteur de vingt énergies est associé à chaque classe. Les énergies sont

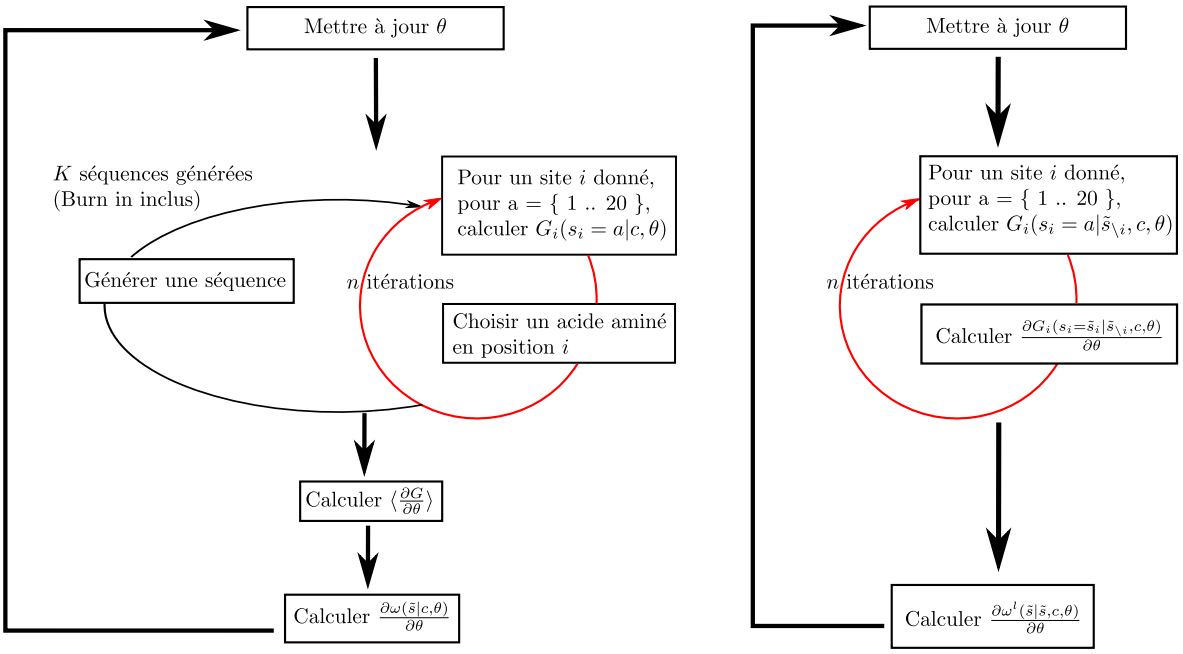


FIGURE 5.7 – Diagramme comparatif des deux algorithmes d'optimisation.

conjointement estimées à l'aide de l'approximation *leave-one-out* et représentent l'affinité de chaque acide aminé pour chaque classe. Ce travail a avant tout été poursuivi par C. L. Kleinman [Kleinman et al., Submitted].

De mon côté, je me suis attachée à la formalisation du problème de l'approximation du facteur de normalisation $\ln Z_s$. Dans le modèle développé précédemment (chapitres 4 et 5), tout comme dans le travail développé par C.L. Kleinman, dans l'équation de la formule de Boltzmann :

$$p(c|s, M) = \frac{e^{-E(s,c)/kT}}{Z_s}, \tag{5.31}$$

où

$$Z_s = \sum_{c \in \mathbb{C}} e^{-E(s,c)/kT}, \tag{5.32}$$

le modèle approxime le facteur de normalisation Z_s à l'aide du *random energy model* :

$$Z_s \simeq \sum_i \mu_{s_i}. \tag{5.33}$$

Cependant, ce facteur peut être approximé d'une manière différente, en l'exprimant comme une moyenne sur des structures alternatives, ou *decoys* (chapitre 7).

Entretemps, nous nous sommes aperçus que la formulation Bayésienne du problème par l'équation :

$$p(s|c) \propto p(s)p(c|s), \tag{5.34}$$

telle qu'initialement proposée dans le premier article [Kleinman et al., 2006], présente plusieurs désavantages, et nous avons donc décidé de reformuler entièrement le problème, tout en analysant les implications des différents termes (potentiel, facteur de normalisation Z_s...) dans le modèle d'évolution. Ceci est le sujet du chapitre 6.

Chapitre 6

Reformulation du probleme

6.1 Introduction

Au chapitre 4, nous définissions la probabilité d'observer une séquence sachant la structure comme :

$$p(s|c) \propto p(s) \cdot p(c|s), \tag{6.1}$$

où $p(c|s)$ était la probabilité physique (distribution de Boltzmann dans l'espace des conformations) et $p(s)$ était une prior, pouvant être quelconque, et que nous avions prise uniforme. Cette formulation bayésienne du problème est très séduisante à première vue. Elle s'apparente aux méthodes d'apprentissage statistiques standards. Toutefois, cette formulation présente le désavantage de masquer sous la prior les subtilités du modèle d'évolution, ce qui est contradictoire avec le but de construire un modèle d'évolution mécanistique.

Une autre approche consiste à redéfinir le cadre statistique de la méthode plus directement selon les termes du modèle d'évolution sous-jacent. Cette approche a d'abord été utilisée par Choi et al, et suppose que les protéines présentes dans les bases de données sont à l'équilibre mutation/sélection [Choi et al., 2007, Choi et al., 2008]. Ainsi, pour un modèle d'évolution donné, il est possible de déterminer la probabilité stationnaire des séquences, sachant leur structure (puisque le modèle suppose que cette structure est conservée par les protéines ancestrales et actuelles), $\varphi_{\theta,c}(s)$. Cette probabilité stationnaire dépend évidemment de la structure, mais également de paramètres θ qui englobent les paramètres mutationnels et ceux qui définissent le potentiel statistique. Si le jeu de données est $(\tilde{s}_k, \tilde{c}_k)_{k=1..K}$ alors la vraisemblance est donnée par :

$$\prod_{1 \leq k \leq K} \varphi_{\theta,\tilde{c}_k}(\tilde{s}_k). \tag{6.2}$$

On peut ensuite optimiser les paramètres du potentiel par la méthode du maximum de vraisemblance définie aux chapitres précédents.

On peut cependant remarquer, en guise d'alternative à la méthode du maximum de vraisemblance, qu'un traitement Bayésien du problème serait possible, en assignant une prior sur le jeu de paramètres θ. Mais, en tout état de cause, il est incorrect de définir une prior sur les séquences, comme nous le faisions auparavant.

6.2 Optimizing statistical potentials for SC phylogenetic models

Authors : Cécile Bonnard, Claudia L. Kleinman, Nicolas Rodrigue, Nicolas Lartillot.

6.2.1 Abstract

Background : Several mutation/selection, structurally constrained (SC), models have been proposed since the first neutral model of protein evolution [Bastolla et al., 1999]. They all use statistical potentials to measure the fit between the encoded amino acid sequence and the structure of the protein. The formulation of these potentials entails different aspects of protein folding/protein design [Robinson et al., 2003,Kleinman et al., 2006], mutation-selection equilibrium [Choi et al., 2007] and specificity [Bastolla et al., 2006]. In a previous work, we proposed a likelihood framework especially meant for protein design and SC models. However, as the method did not correctly account for the contribution of the mutation pressure to the composition of the proteins used at the training step, including the resulting potentials in a SC model resulted in counting the mutational component twice.

Results : We present here a reformulation of our statistical framework, in terms of mutation/selection equilibrium, and we show how to include a correction for mutation in the potentials before using them in a SC context. To illustrate this point, we compare the influence of the different components (specificity and mutation-correction) using the Miyazawa and Jernigan's potential and the potential we previously optimized [Bonnard et al., 2009].

Conclusions : We point out that a mutation-correction is needed by the potentials when we include them into a SC phylogenetic model, and that the SC model using mutation-corrected potentials are better that those using non-corrected potentials.

6.2.2 Background

Mutation/selection models based on a structural constraint (SC) [Bastolla et al., 1999, Robinson et al., 2003,Choi et al., 2008,Rodrigue et al., 2009] are attractive. In these models formulated at the codon level, the mutation process is defined at the nucleotide level, while selection is applied to the encoded amino acid sequence, via a scoring function, measuring how the sequence fits the structure. A first purely neutral evolution model was presented in [Bastolla et al., 1999]. The authors presented the structurally constrained neutral (SCN) model, in which mutations that are not purely neutral are selected against. A mutation is considered as neutral if it does not change the thermodynamic stability of the protein, which was evaluated by existing statistical potentials [Miyazawa and Jernigan, 1985]. This model was first used to show that biological sequences evolution can be seen as a random walk on a neutral network, caused by random drift. The model was further refined using improved statistical potentials [Bastolla et al., 2000] to show that such a model gives profiles of amino acid propensities along the sequences that correlate well with empirical sequences in PDB. In 2003, [Robinson et al., 2003] used Monte Carlo Markov Chain (MCMC) methods to make empirical inference about the evolutionary process separating two sequences, using preexisting statistical potentials [Jones et al., 1992a]. This approach was then generalized to more than two sequences by [Rodrigue et al., 2005], with different statistical potentials [Miyazawa and Jernigan, 1985,Kleinman et al., 2006]. Then, Choi [Choi et al., 2007, Choi et al., 2008], assuming that proteins are at the mutation/selection equilibrium, assessed the relationship between structural constraints and observed sequences for proteins from the PDB. Altogether, these previous developments define the broad lines of a powerful framework to quantify the relative influence of mutation, selection and genetic drift in protein evolution.

However, in these approaches, which all use statistical potentials, there may be a problem concerning the definition of the thermodynamic stability. Indeed, the sequence has to fold into the structure, but has also to not fold into other alternative structures. If the potential presented by [Bastolla et al., 1999] is made so as to maximize the contrast between native and alternative structures, others approaches do not explicitly take misfolding into account, e.g. [Robinson et al., 2003]. If this problem is important in protein folding, it is probably also relevant in a protein design approach. Another potential problem is the adequacy of currently available statistical potentials, with respect to a folding/inverse folding issue.

Structures evolve slowly [Chothia and Lesk, 1986] compared to sequences, and in all models mentioned above, it has been considered as fixed throughout evolution. In this

context, evolution can be seen as an algorithm performing protein design (i.e. finding the sequences that fit the structure of interest), while accommodating some amount of noise due to random drift. Protein design entails a search both in sequence space and in structure space, so as to find sequences having the fixed structure as their ground state. However, most statistical potentials currently available have been devised in a context of protein folding, and it has been argued that such potentials might not be optimal in a protein design context [Chiu and Goldstein, 1998a, Rossi et al., 2001]. In part for that reason, and also to establish a more systematic basis for statistical testing and empirical fitting of structurally constrained models, we previously proposed a maximum likelihood framework to optimize statistical potentials especially meant for protein design [Kleinman et al., 2006].

However, in this previous work, although we did address the specificity issue, we did not correctly formalize the complex interplay of mutation and selection. Specifically, our likelihood was defined in terms of observing the sequence of the database ($\tilde{s}$) given the conformation ($\tilde{c}$) and the parameters of the potential (θ) :

$$L(\theta) = p(\tilde{s}|\tilde{c}, \theta). \tag{6.3}$$

This probability was obtained by applying Bayes inversion formula :

$$p(\tilde{s}|\tilde{c}, \theta) \propto p(\tilde{c}|\tilde{s}, \theta) \cdot p(\tilde{s}), \tag{6.4}$$

where

$$p(\tilde{c}|\tilde{s}, \theta) = \frac{1}{Z_s} \cdot e^{-\beta E(\tilde{s}, \tilde{c})}, \tag{6.5}$$

and Z_s is a normalization factor. This normalization factor is a sum over all conformations, and therefore, through this factor, the misfolding problem was accounted for : the native sequence-structure couple has to be more stable than the sequence folded into alternative structures. However, the fact that $p(s)$ was loosely defined as a uniform prior implied that the mutational input was entirely neglected.

Choi et al [Choi et al., 2007], in contrast, did correctly pose the problem in term of a mutation/selection equilibrium. On the other hand, they did not address the specificity problem. Borrowing from [Choi et al., 2007], we reformulate here our statistical framework so as to correctly account for specificity and protein design and include a corrected term for mutation aspects. First, we assume that the proteins are at the mutation-selection equilibrium, and we thus redefine the probability of finding a sequence in a structure in our maximum likelihood framework. Then, this new definition allows us to include the dependence to misfolded structures in the model. To illustrate the empirical impact of each

issue raised above, we use the well-known Miyazawa and Jernigan's potential [Miyazawa and Jernigan, 1985], and analyze the improvement brought by each approach. Then, we accordingly modified our previous potential, and show that the overall reformulation results in a better fit than we previously showed.

6.2.3 Results

6.2.3.1 Evolutionary model

Let σ and σ' be two nucleotide sequences differing by only one position, and s (resp. s') the amino acid sequence encoded by σ (resp. σ'). The key evolutionary process defined in a SC model can be expressed as :

$$R_{\sigma\sigma'} = Q^{mut}_{\sigma\sigma'} e^{\frac{\beta}{2}\left(H(s,c|\theta)-H(s',c|\theta)\right)}, \tag{6.6}$$

where $R_{\sigma\sigma'}$ is the substitution rate between σ and σ, $Q^{mut}_{\sigma\sigma'}$ is the mutation matrix between σ and σ', and β stands for the global strength of selection. $H(s,c|\theta)$ is a function depending on the amino acid sequence and on the three-dimensional protein structure c (which remains constant along the evolutionary tree) and θ is the set of parameters of the scoring function. $R_{\sigma\sigma'}$ is a $4^{3n} \cdot 4^{3n}$ matrix which is equal to zero if σ and σ' differ by more than one nucleotide, or if the mutation leads to a stop codon in σ'. Note that if $\beta = 0$, the model is the pure mutation model (except for purification of stop codons), called M_0. We do not take the mutations leading to stop codons (TAA, TAG, TGA) into account because they are considered as too detrimental for the sequence.

The evolutionary model described by eq. (6.6) is a reversible Markov process whose stationary distribution, given a set of parameters Θ (which include θ) and a structure c, is :

$$\varphi_{\Theta,c}(\sigma) = \frac{1}{Y_c} \cdot \Pi^{mut}_{\sigma} \cdot e^{-\beta H(s,c|\theta)}, \tag{6.7}$$

where

$$Y_c = \sum_{\sigma'} \Pi^{mut}_{\sigma'} \cdot e^{-\beta H(s',c|\theta)} \tag{6.8}$$

is a normalization factor, and Π^{mut}_{σ} is the stationary distribution of the mutation process. Specifically, Π^{mut}_{σ} is the product over all positions i in σ of the stationary distributions of the nucleotide at this position (σ_i) :

$$\Pi^{mut}_{\sigma} = \frac{\prod_{1\leq i\leq 3n} \pi_{\sigma_i}}{1-\pi_{stop}}, \tag{6.9}$$

where $\pi_{stop} = \pi_{TTA} + \pi_{TAG} + \pi_{TGA}$.

One of the main problems of a SC model is the description of the function $H(s, c|\theta)$ and of its parameters. A full likelihood model entirely defined in a phylogenetic framework would be the ideal approach. However, it would be too computationally demanding, as it would ideally require a dataset of aligned amino acid sequences, and their associated nucleotide sequences and phylogenies. Instead, we optimize the parameters of the scoring function $H(s, c|\theta)$ in an independent procedure. However, we still want this optimization to be consistent with the underlying evolutionary model. To this aim, we first obtain the stationary distribution of an amino acid sequence s :

$$\varphi_{\Theta,c}(s) = \sum_{\sigma|s} \varphi_{\Theta,c}(\sigma), \tag{6.10}$$

where the sum is over all the nucleotide sequences σ coding for a sequence s. Thus, we can write :

$$\varphi_{\Theta,c}(s) = \sum_{\sigma|s} \frac{1}{Y_c} \cdot \Pi_{\sigma}^{mut} e^{-\beta(H(s,c|\theta))} \tag{6.11}$$

$$= \Pi_{s}^{mut} \frac{1}{Y_c} \cdot e^{-\beta(H(s,c|\theta))} \tag{6.12}$$

$$= \frac{1}{Y_c} \cdot e^{-\beta\left(H(s,c|\theta)+\sum_{1\leq i\leq n} \nu_{s_i}^{mut}\right)}, \tag{6.13}$$

where

$$\nu_{s_i}^{mut} = \sum_{\sigma_1\sigma_2\sigma_3|s_i} -\ln \pi_{\sigma_1}^{mut} \pi_{\sigma_2}^{mut} \pi_{\sigma_3}^{mut}. \tag{6.14}$$

which can be seen as 'mutational free energies'.

6.2.3.2 Definition of the scoring function

In this article, we chose to define the scoring function as a function of the probability $f(\tilde{c}|s)$, defined in a protein design context [Seno et al., 1998].

From the laws of thermodynamics, we can express the probability of having a fixed amino acid sequence s in a three-dimensional structure c as a Boltzmann distribution :

$$f(c|s) = \frac{\exp\left(-\beta E(s,c)\right)}{Z_s}, \tag{6.15}$$

where

$$Z_s = \sum_{c' \in \mathbb{C}} \exp\left(-\beta E(s,c')\right) \tag{6.16}$$

is a normalization factor, constant for a given sequence and $E(s,c)$ is the energy of s folded into c, and $\beta = 1/kT$ is the inverse temperature (not to be confused with the strength of selection in equation 4). For notation simplicity, we do not represent the dependance to the parameters of the energy function, which are included in θ. Eq. (6.15) is the distribution probability used in protein folding approaches, which aims at finding the *ground state* $\hat{c}$ of the sequence s such as :

$$f(\hat{c}|s) \geq f(c|s) \ \forall c. \tag{6.17}$$

We assume that the native structure of s, $\tilde{c}$, is its ground state, for the chosen energy function. Biological adaptation would ensure that this ground state is well defined (i.e. $p(\tilde{c}|s)$ is close to 1), although this may not be true for all proteins.

The aim of protein design methods is to find a sequence which actually folds into a given native structure. Let $\tilde{c}$ be the structure of interest. The probability of having the sequence s in the conformation $\tilde{c}$ is thus equal to $f(\tilde{c}|s)$ [Seno et al., 1996]. Then, we want to find a sequence $\hat{s}$ so that :

$$f(\tilde{c}|\hat{s}) \geq f(\tilde{c}|s) \ \forall s. \tag{6.18}$$

The parallel between SC model and protein design made in the introduction suggest that we can define $H^{SC}(s,\tilde{c}) = -\ln f(\tilde{c}|s) = E(s,\tilde{c}) + \ln Z_s$.

We can expand $\ln Z_s$ in powers of β :

$$\ln Z_s(\beta_1) = \ln Z_s(\beta_0) + \beta \frac{\partial \ln Z_s}{\partial \beta_1}(\beta_0) + \cdots \ . \tag{6.19}$$

As

$$\frac{\partial \ln Z_s}{\partial \beta_1} = -\sum_{c' \in \mathbb{C}} \frac{E(s,c') e^{\left(-\beta_1 E(s,c')\right)}}{\sum_{c'' \in \mathbb{C}} e^{\left(-\beta_1 E(s,c'')\right)}} = -\sum_{c' \in \mathbb{C}} E(s,c') \cdot f(c'|s), \tag{6.20}$$

then, if $\beta_0 = \beta$ and $\beta_1 = 0$:

$$\ln Z_s(0) \simeq \ln Z_s(\beta) - \beta \sum_{c' \in \mathbb{C}} E(s,c') \cdot f(c'|s) \tag{6.21}$$

$$\ln Z_s(\beta) \simeq \ln Z_s(0) + \beta \sum_{c' \in \mathbb{C}} E(s,c') \cdot f(c'|s) \tag{6.22}$$

$$\ln Z_s(\beta) \simeq \ln Z_s(0) + \beta \langle E(s,c') \rangle_{c' \in \mathbb{C}}. \tag{6.23}$$

Thus considering that the rest of the development of $\ln Z_s$ is negligible, the main contribution to $\ln Z_s$ is the average energy of structures weighted by their Boltzmann probability $p(c'|s)$ (eq. 6.15). Different proposals can be made to approximate $\langle E(s,c') \rangle_{c' \in \mathbb{C}}$, e.g. minimizing a Z-score between the native and alternative structures [Mirny and Shakhnovich,

1996], or minimizing the energy of the native structure compared to average energy of alternative structures [Deutsch and Kurowski, 1996]. However, to simplify the argument, we choose the simplest approximation based on the random energy model (REM) [Shakhnovich and Gutin, 1993, Finkelstein, 1997] :

$$\ln Z_s = \sum_{1 \leq i \leq n} \lambda_{s_i}, \tag{6.24}$$

where the parameters λ_a are analogous to chemical potentials. In this model, we assume that the average $\langle E(s, c') \rangle_{c' \in \mathbb{C}}$ does not depend on the specific order on which the amino acids appears in s, but only on the global composition.

6.2.3.3 Mutation-selection equilibrium

If we suppose that the natural proteins are at the mutation-selection equilibrium, we can write the probability of a sequence with a fixed structure, for a set of p natural proteins (structure-sequence pairs (s, c)) taken from the PDB as :

$$L(\theta) = P(s^p | c^p, \theta) = \prod_p \varphi_{\Theta, c^p}(s^p, \theta). \tag{6.25}$$

This can be seen as a likelihood, to be optimized with respect to the parameters Θ. For simplicity, the method will be described for a single protein, but it can be easily generalized to the whole set of proteins. Thus, for a natural protein taken from the PDB, combining eq. (6.13) and (6.25) :

$$P(s|c, \Theta) = \sum_{\sigma | s} \frac{1}{Y_c} \cdot e^{-\beta \left(H(s,c|\theta) + \sum_{1 \leq i \leq n} \nu^{mut}_{s_i} \right)}. \tag{6.26}$$

Under the REM model, this equation becomes :

$$P(s|c, \theta) \propto \exp -\beta \Big(\underbrace{E(s, c) + \sum_{1 \leq i \leq n} \lambda_{s_i}}_{H(s,c|\theta)} + \sum_{1 \leq i \leq n} \nu_{s_i} \Big). \tag{6.27}$$

If the training was done using nucleotide sequences, we could in principle separately estimate the λ_a and the ν_a. In the present case, however, the model is trained directly on amino-acid sequences. The two parameter vectors λ_a and ν_a can then be merged into one single vector μ_a :

$$\mu_a = \lambda_a + \nu_a. \tag{6.28}$$

As a result, the ν^{mut}_a parameters, corresponding to the contribution of the mutation pressure to the mutation/selection equilibrium, are embedded in the parameters of $H(s, c|\theta)$.

This is not a problem, as long as we take care of removing them when we use the potential in the phylogenetic context. To obtain this mutation-corrected potential, one has to retrieve the mutation component from the nucleotide sequences taken from the learning database (using protogene [Moretti et al., 2006]) and then estimate the mutational free energies (ν_a^{mut}) based on the frequencies of four nucleotides for four-fold degenerate sites. Finally, we have to correct the potential by subtracting these mutational free energies, before integrating the corrected potential into the phylogenetic model. In previous works [Kleinman et al., 2006, Bonnard et al., 2009], we did not perform this correction.

6.2.3.4 Studied models

To evaluate the impact of the problems mentioned above (accounting for specificity, and correcting for the mutation pressure), we first implement each combination using the Miyazawa and Jernigan's potential (MJ) [Miyazawa and Jernigan, 1985].

Let us define $E^{MJ}(s,c)$ as the energy of the sequence s in the conformation c using the MJ potential :

$$E^{MJ}(s,c) = \sum_{1 \leq i < j \leq N} \Delta_{ij} \varepsilon^{MJ}_{s_i s_j}, \tag{6.29}$$

where Δ_{ij} represent the contact map of the structure c, and $\Delta_{ij} = 1$ if the sites i and j are distant by less than a contact threshold. ε^{MJ}_{ab} are the contact potentials between the two amino acids a and b defined by [Miyazawa and Jernigan, 1985]. These contact potentials have been estimated on a database, using a quasi-chemical approximation [Miyazawa and Jernigan, 1985].

First, we will consider the impact of the pure MJ potential to the SC model, MJ^0_β :

$$R^{MJ0}_{\sigma\sigma'} = Q^{mut}_{\sigma\sigma'} e^{\frac{\beta}{2}\left(E^{MJ}(s,c) - E^{MJ}(s',c)\right)}. \tag{6.30}$$

This is similar to what had been done by [Robinson et al., 2003, Choi et al., 2007, Choi et al., 2008]. However, as we saw in 6.2.3.2, this does not account for specificity. Thus, we include in the MJ^{REM}_β model the term corresponding to the random energy approximation of $\ln Z_s$ defined by eq(6.24) :

$$R^{MJ_{REM}}_{\sigma\sigma'} = Q^{mut}_{\sigma\sigma'} e^{\frac{\beta}{2}\left(H^{MJ}(s,c|\theta) - H^{MJ}(s',c|\theta)\right)}, \tag{6.31}$$

where

$$H^{MJ}(s,c|\theta) = E^{MJ}(s,c) + \sum_{1 \leq i \leq n} \lambda_{s_i}, \tag{6.32}$$

where the λ_a have been optimized using the method of [Kleinman et al., 2006], on the dataset DSL introduced in [Bonnard et al., 2009].

However, this model includes an overestimation of the mutational pressure, which is counted twice : once in the Q^{mut} matrix, and once in $H^{MJ}(s, c|\theta)$. To correct this problem, we make the retrospective correction explained above (eq. (6.26)), leading to the MJ_β^{cor} model defined by :

$$R_{\sigma\sigma'}^{MJcor} = Q_{\sigma\sigma'}^{mut} e^{\frac{\beta}{2}\left(H^{MJ}(s,c|\theta)+\sum_{1\leq i\leq n}\nu_{s_i}^{mut}-\left(H^{MJ}(s',c|\theta)+\sum_{1\leq i\leq n}\nu_{s'_i}^{mut}\right)\right)}, \tag{6.33}$$

with $H^{MJ}(s, c|\theta)$ defined in eq. (6.32). Through these three models, we aim at characterizing the differences in the fit of the models, induced by each approximation.

Finally, we compare our previous model presented in [Kleinman et al., 2006, Bonnard et al., 2009] to the corrected model. We first define the potential :

$$E^{pot}(s, c) = \sum_{1\leq i<j\leq n} \Delta_{ij}\varepsilon_{s_i s_j} + \sum_{1\leq i\leq n} \alpha_a^{B_i}, \tag{6.34}$$

where ε_{ab} is the contact potential between the two amino acid a and b. B_i represents the solvent accessibility of site i in structure c and α_a^d represents the solvent accessibility potential of a placed in the solvent accessibility class $d = 1..D$ (D is the number of solvent accessibility classes). Δ_{ij} and B_i are defined by the structure c, while ε_{ab} and α_a^d are parameters that have to be optimized under the following constraints [Kleinman et al., 2006] :

$$\sum_{1\leq a\leq 20} \alpha_a^d = 0,\ d = 1..D, \tag{6.35}$$

$$\sum_{1\leq a\leq 20}\sum_{1\leq b\leq 20} \varepsilon_{ab} = 0. \tag{6.36}$$

Note also that the μ_a are embedded in the solvent accessibility parameters [Bonnard et al., 2009].

Our previous model SC_β^{init} was defined as :

$$R_{\sigma\sigma'}^{init} = Q_{\sigma\sigma'}^{mut} e^{\frac{\beta}{2}\left(H^{pot}(s,c|\theta)-H^{pot}(s',c|\theta)\right)}, \tag{6.37}$$

where

$$H^{pot}(s, c|\theta) = E^{pot}(s, c|\theta) + \sum_{1\leq i\leq n} \lambda_{s_i}, \tag{6.38}$$

where $E^{pot}(s, c|\theta)$ is the statistical potential defined in our previous articles [Kleinman et al., 2006, Bonnard et al., 2009].

This model is compared to the mutation-corrected model, SC_β^{cor}, whose substitution process is defined by :

$$R_{\sigma\sigma'}^{cor} = Q_{\sigma\sigma'}^{mut} e^{\frac{\beta}{2}\left(H^{pot}(s,c|\theta)+\sum_{1\leq i\leq n}\nu_{s_i}^{mut}-\left(H^{pot}(s',c|\theta)+\sum_{1\leq i\leq n}\nu_{s'_i}^{mut}\right)\right)}. \quad (6.39)$$

6.2.4 Phylogenetic analysis

The comparison of the models is performed by the computation of the Bayes factor (see Methods, [Rodrigue et al., 2006, Bonnard et al., 2009]). A Bayes factor greater (resp. smaller) than 1 is considered as an evidence in favor of (resp. against) the model M_1. In table 6.1, we present the maximum Bayes factor obtained using the MJ potential and the terms that each model includes in its definition : $E(s,c)$ for the inclusion of a statistical potential, $\ln Z_s$ if it includes an estimate of the normalization term, and ν^{mut} if the retrospective mutation-correction is made (which is only needed if we take $\ln Z_s$ into account). The first model, MJ_β^0 includes only the statistical potential of Miyazawa

		GLOBIN	$E(s,c)$	$\ln Z_s$	ν^{mut}
MJ_β^0	BF	[46.505 :46.967]	+	-	-
	β	[0.148 :0.172]			
MJ_β^{REM}	BF	[68.224 :69.758]	+	+	!
	β	[0.294 :0.296]			
MJ_β^{cor}	BF	[89.877 :90.061]	+	+	+
	β	[0.354 :0.36]			

TABLE 6.1 – BF : maximum natural logarithm of the Bayes factor, β : optimal β under which these values were obtained.

and Jernigan. The second model, MJ_β^{REM} includes the MJ statistical potential and $\ln Z_s$ (estimated on dataset DSL) but potentially overestimates the mutational term. The last model includes all terms : the MJ statistical potential, the normalization term and the retrospective correction of the mutational term (cf. Methods). The inclusion of $\ln Z_s$ into the model yields better Bayes factor than the simple MJ_β^0 model for the dataset GLOBIN. However the careless inclusion of an estimate of $\ln Z_s$, including a mutational component, may lead to a suboptimal model. Indeed, table 6.1 shows that the inclusion of the retrospective correction of the mutation term gives a better Bayes factor than simply including the normalization factor for our dataset.

Next, we perform a similar comparison on a potential previously developed, and optimized in a protein design context [Kleinman et al., 2006] The results are summarized in table 6.2.

		GLOBIN	$E(s,c)$	$\ln Z_s$	ν^{mut}
SC_β^{init}	BF β	[72.485 :72.872] [0.9 :0.996]	+	+	!
SC_β^{cor}	BF β	[118.779 :119.221] [1.036 :1.056]	+	+	+

TABLE 6.2 – BF : maximum natural logarithm of the Bayes factor, β : optimal β under which these values where obtained.

First, we observe the protein design potential has a globally better fit than the pairwise potential of Miyazawa and Jernigan. In addition, the retrospective correction of the mutational term seems to give a better fit to the SC model, which confirms that the mutation-correction term is needed to keep consistency between the optimization procedure and the SC model.

6.2.5 Conclusions

We described here a reformulation of a full likelihood framework to optimize statistical potentials explicitly suited for phylogenetic SC models. In the model presented here, we optimize a statistical potential in a protein design approach, taking into account the normalization factor Z_s, and we defined a retrospective correction term to correct the double counting of the mutational pressure. We applied the model to a SC model using the statistical potential of Miyazawa and Jernigan to show how various terms ($\ln Z_s$, mutation-correction) can affect the fit under a well-known potential, and then applied it to our potential defined in [Kleinman et al., 2006, Bonnard et al., 2009]. We show that both the Z_s and the correction term are important when we optimize statistical potential for a SC model.

Here, we only used an approximation of the normalization term, inspired from the Random Energy Model, but this is a simple approximation of the structure space, and we are currently testing our model using explicit alternative structures.

6.2.6 Methods

6.2.6.1 Datasets

We used the same optimization dataset as in [Bonnard et al., 2009] : *DSL* (3,363 proteins and 835,717 sites) is made of proteins culled from the PDB with less than 25 % of sequence identity [Wang and Dunbrack, 2003]. The contact maps were defined using a contact threshold of 6.5 Å, and we used Naccess 2.1 [Hubbard and Thornton, 1993] to define the solvent accessible area of each amino acid of each protein. We then partitioned this space into 14 classes [Kleinman et al., 2006].

We used the phylogenetic dataset GLOBIN, which is made of 15 vertebrate sequences of the β-globin gene (taken from the dataset described in [Yang et al., 2000a]) with a tree topology estimated using Phylobayes 3.1c [Lartillot et al., 2009] and with the protein structure defined by the PDB file 4HHB.

To compute the mutational frequency of each nucleotide bases of the learning database, we first retrieve the nucleotide sequences using Protogene [Moretti et al., 2006]. We then count the frequency of each one of the four bases at four-fold degenerate sites.

6.2.6.2 Leave-one-out approximation

The normalization factor Y_c cannot be obtained analytically, because it is a sum over the sequence space (20^n). Two approaches can thus be proposed. The first is to use MCMC, and sample sequences according to the actual probability distribution $P(s|\tilde{c})$. This sampling allows one to compute the gradient of the likelihood [Kleinman et al., 2006]. In [Bonnard et al., 2009], instead of optimizing the true likelihood specified in eq. (6.26), we define a leave-one-out pseudo-likelihood whose normalization factor Y'_c is analytically computable. Although this method is formally dependent on the native sequence, the results obtained using the two approximations are indistinguishable in practice, but the leave-one-out method has an up to a 1,000 fold decrease in computational time [Bonnard et al., 2009].

6.2.6.3 Bayes factor

The Bayes factor is a powerful tool for comparing models. Given the *evidence* (marginal likelihood) of two models M_0 (resp. M_1), $P(A|M_0,\Theta)$ (resp. $P(A|M_1,\Theta)$), the Bayes factor is defined as :

$$B(M_0,M_1|\Theta) = \frac{P(A|M_1,\Theta)}{P(A|M_0,\Theta)}, \tag{6.40}$$

where A stands for the data (alignment) and $P(A|M,\Theta)$ is computed by thermodynamic integration, described in [Rodrigue et al., 2006]. It consists in sampling along the path between the two models through slight variations of Θ.

6.3 Conclusion

La reformulation de la méthode d'optimisation de potentiels statistiques proposée dans cet article, faisant intervenir un modèle mutation/sélection, a permis notamment de mettre en évidence qu'il y avait une redondance dans le modèle au sujet de la pression mutationnelle. En effet, lorsque l'on optimise des potentiels, il nous est impossible de séparer la part mutationnelle dissimulée et inextricablement incluse dans les potentiels chimiques. Par contre, il est facile de la soustraire au potentiel chimique lorsque l'on s'intéresse au modèle d'évolution.

A l'aide de l'algorithme d'optimisation défini au chapitre 5 et de la formulation mathématique du contexte définie dans ce chapitre, il devient facile d'étudier de nouvelles formulations du facteur de normalisation Z_s, qui correspond à la somme des énergies de la séquence repliée dans l'ensemble des structures possibles pour la séquence considérée :

$$Z_s = \sum_{c \in \mathbb{C}} e^{-E(s,c)}. \tag{6.41}$$

Les approximations utilisées précédemment pour le $\ln Z_s$ étaient liées au *random energy model*, qui est l'approximation la plus simple pour ce facteur de normalisation [Shakhnovich and Gutin, 1993]. D'autres potentiels statistiques, et en particulier ceux présentés dans [Bastolla et al., 2000, Bastolla et al., 2006], ont été optimisés afin de maximiser de manière explicite la différence d'énergie entre la protéine native et la séquence native repliée dans des structures leurres. Ces potentiels sont donc construits de manière à ce que la séquence native préfère la structure native aux structures alternatives. Le but de la dernière partie de cette thèse et de développer un cadre méthodologique afin d'intégrer des structures leurres à la fois dans la méthode d'optimisation des potentiels (toujours dans un contexte de *protein design*) mais aussi dans le modèle d'évolution soumis à des contraintes structurales.

Chapitre 7

Inclusion de structures leurres

7.1 Introduction

Dans les chapitres précédents, nous avons déterminé une méthode d'optimisation de potentiels statistiques, pour les intégrer dans un modèle d'évolution soumis à des contraintes structurales. Cependant, si de nouvelles formes de potentiel ont été explorées à l'aide de cette méthode [Kleinman et al., Submitted], celle-ci fait appel à une approximation simple d'un facteur de normalisation important, Z_s, basé sur le *random energy model*. Spécifiquement, la probabilité à maximiser, $p(c|s)$ est exprimée de la manière suivante :

$$p(c|s) = \frac{e^{-E(s,c)/kT}}{\sum_{c \in \mathbb{C}} e^{-E(s,c)/kT}} = \frac{e^{-E(s,c)/kT}}{Z_s}, \tag{7.1}$$

où Z_s était approximé par :

$$\ln Z_s \simeq \sum_i \mu_{s_i}, \tag{7.2}$$

où s_i est l'acide aminé présent au site i. Or cette approximation est couramment critiquée car elle ne prend pas en compte des structures réellement existantes.

En utilisant la reformulation du modèle présentée au chapitre 6, et de l'approximation proposée dans [Deutsch and Kurowski, 1996], il devient possible d'intégrer directement des structures leurres dans l'approximation de $\ln Z_s$, en proposant que, à une constante additive près :

$$\ln Z_s \simeq \langle -E(s, c') \rangle_{c' \in \mathbb{C}}, \tag{7.3}$$

où $\mathbb{C}$ est un ensemble de structures leurres adéquat. La linéarité de l'équation (7.3) fait que le $\ln Z_s$ approximé ainsi peut être facilement intégré dans le potentiel, et donc dans la méthode d'optimisation.

La problématique devient alors de trouver un bon jeu de structures permettant d'approximer cet espace $\mathbb{C}$. Deux cas de figure peuvent se présenter : on peut choisir un jeu de structures représentant l'ensemble réel des structures (en construisant par exemple un jeu de structures par *threading*), ou considérer que les structures qui doivent être préférentiellement écartées par le potentiel sont des structures les plus compétitives par rapport à la structure native.

7.2 Inclusion of decoys in the optimization of a statistical potential made for phylogenetic models

Authors : Cécile Bonnard, Nicolas Rodrigue, Claudia L. Kleinman, and Nicolas Lartillot.

7.2.1 Abstract

Background : Structurally constrained (SC) models have been proposed in order to account for the fact that the evolution of a protein sequence can be partly expressed as a function of its structure. In such models, the connection between structure and sequence is made by statistical potentials which provide a quantification of the compatibility of a sequence with a given structure. In previous articles, we defined an optimization procedure of statistical potentials especially meant for protein design and for phylogenetic SC models. Two approaches can be proposed to optimize statistical potentials : on one hand, the native protein must have the lowest energy (positive design) and on the other hand, the misfolded, competitive structures must be ruled out (negative design). However, in our previous work, we only performed positive design. Here, we modify the framework we previously defined to take into account misfolded structures (decoys).

Results : We tested two different structure descriptions (a contact potential and a distance potential) and different decoy datasets (threading, high-resolution decoy, and variational decoys), and compare the results obtained with those from our previous model. Cross validation scores of the different methods show that the inclusion of decoys gives slightly better results under the contact and the distance potentials. In addition, Bayes Factor evaluations in the phylogenetic SC cortex seems to favor the inclusion of decoys.

Conclusions : The statistical optimization framework presented here results in significant improvements to the SC model. On the other hand, the relatively modest improvement suggests that the decoy datasets obtained by threading are not adapted to the current purpose. Further

investigations are thus needed, in order to test new forms of decoy datasets and variational decoys.

7.2.2 Background

The structure and the sequence of proteins are deeply interconnected during evolution. As a general rule, protein sequences seem to evolve rapidly, compared to the three-dimensional structure. For instance, mammalian hemoglobins display substantial sequence variation [Chothia and Lesk, 1986] while their three-dimensional structures are virtually identical. This observation motivates the use of a structure-dependent quasi-neutral process for describing protein sequence evolution. In this model, a fixed structure is considered, and is assumed to be a constraint on sequence evolution. Substitutions destabilizing the conformation are then selected against, whereas those having a much less dramatic effect on conformational stability pass through the sieve of natural selection.

Over recent years, attempts have been made to formalize and implement this idea of quasi-neutral, conformation-dependent evolution of protein sequences, in what have been called structurally constrained (SC) phylogenetic models. They are mostly formulated at the level of the codon sequence, and propose that the rate of substitution is the product of the mutation rate (defined at the nucleotide level), and a selection factor (defined at the level of the encoded amino acid sequence) meant as a proxy for the ratio between the fixation probability under the current model and the neutral fixation probability [Robinson et al., 2003, Rodrigue et al., 2009] :

$$R_{\sigma\sigma'} = Q^{mut}_{\sigma\sigma'} e^{\frac{\beta}{2}(H(s,c)-H(s',c))}. \tag{7.4}$$

Here, $Q^{mut}_{\sigma\sigma'}$ is the mutation rate from the nucleotide sequence σ to the sequence σ', and $H(s,c)$ is a score function which measures how well amino acid sequence s (expected by the nucleotide sequence σ) fits the structure c.

For general scoring functions, SC models imply that the substitution process at a given site is dependent on the entire sequence of the protein at the evolutionary instant being considered, and therefore, standard dynamic programming algorithms [Felsenstein, 1981] for computing the likelihood cannot be used. Instead, Bayesian Monte Carlo methods have been proposed [Robinson et al., 2003, Rodrigue et al., 2005], based on the explicit sampling of the detailed substitution history along the branches of the phylogeny. Using Bayes factors, SC models can be compared with other phylogenetic models [Rodrigue et al., 2009], and with each other [Rodrigue et al., 2009]. In particular, Rodrigue et al [Rodrigue et al., 2009] showed that SC models entail a delicate tradeoff between two conflicting needs. On

one hand, the scoring functions currently used in these models are still too simple to outperform more phenomenologically motivated codon models based on the direct estimation of the ratio of non-synonymous over synonymous substitutions, e.g. [Yang et al., 2000a], which clearly suggests that more sophisticated models of sequence-structure compatibility should be used. On the other hand, as the scoring function has to be computed for each possible mutation, and for each substitution event along the phylogenetic tree, this computation has to be inexpensive. Thus far, this trade-off has been handled by using statistical potentials, which are simplified scoring functions mimicking a physical conformational (free) energy. In contrast to physical force fields, however, statistical potentials are not derived from first principles, but using statistical learning methods. Our aim is therefore to create statistical potentials especially suited to SC models.

In a previous article, we defined a maximum likelihood framework to optimize the parameters of a simple pseudo-energy function using databases of sequence-structure pairs [Kleinman et al., 2006]. We then proposed a computationally improved optimization method, based on a leave-one-out pseudo-likelihood. We showed that, compared to the strict likelihood framework developed previously, the leave-one-out method provides equivalent parameters, although with a considerable computational gain [Bonnard et al., 2009]. The resulting computational breakthrough allows more extensive empirical testing of alternative forms of statistical potentials, including parameters representing torsion angles along the main chain, secondary structure determinants, or distance-dependent pairwise energy terms [Kleinman et al., Submitted].

However, the parameters of the potentials introduced in [Kleinman et al., 2006] are optimized so that the native protein (sequence-structure pair) has the lowest energy (positive design). Yet, it is widely accepted that the native sequence in its native structure does not only have the lowest energy, but also has a lower energy than the same sequence in alternative misfolded structures. In this article, we introduce an optimization method based on the explicit use of competitive, misfolded structures (or *decoys* for short), in a so-called negative design approach.

Misfolded structures have been included in pseudo energy functions early on using a variety of methods [Deutsch and Kurowski, 1996, Seno et al., 1998, Bastolla et al., 2000]. Deutsch and Kurosky maximized the gap between the energy of the effective structure and the average energy over all the decoys [Deutsch and Kurowski, 1996]. Chiu and Goldstein define a modified Z-score, which represents the divergence between misfolded and native structure, and propose to find the sequence [Chiu and Goldstein, 1998a] or the structure [Chiu and Goldstein, 1998b] which maximize the Z-score. Bastolla et al [Bastolla et al., 2006] relied on a measure of the average overlap between the native sequence and

misfolded structures, to derive a statistical potential. This potential was then used as a score in a stochastic model of protein evolution [Bastolla et al., 2006] and represents the first attempt to include decoys in a structurally constrained model.

The inclusion of misfolded structures into the scoring scheme raises the question of how to properly define the set of competing structures. There now exist different methods to construct decoy datasets. Most of the decoy datasets are obtained by threading the native sequence into alternative (real) structures taken from the PDB. This leads to gapless (where the sequence is folded into a whole unique structure) and gapped threading (where the sequence is folded into parts of several structures which are then artificially linked together) [Finkelstein, 1997]. Another method consists in determining the exhaustive set of near-native structures which can be defined in a lattice model, and then refine the dataset [Rajgaria et al., 2008].

Given the wide range of possible approaches for accounting for specificity and for defining decoys, we decided to set up a statistical framework for comparing alternative models based on decoys. Using a simple form of statistical potential [Kleinman et al., 2006, Bonnard et al., 2009], we show that the inclusion of decoys into the potential optimization procedure leads to an improvement of the fit of the model as measured by cross validation in a protein design settings, or by Bayes factors evaluation in a phylogenetic context. This improvement, however, is modest and is highly dependent on the choice of the decoy set. We also define a variational method, based on *virtual decoys*, which globally outperforms all other settings based on explicit decoy sets. Altogether, our results are promising, while suggesting that improvements on currently existing decoy sets are still needed.

7.2.3 Methods

7.2.3.1 Datasets

The original dataset was taken from [Rajgaria et al., 2008] and was culled using PISCES [Wang and Dunbrack, 2003] to make a dataset DS of 583 proteins (64,558 sites), with less than 25 % sequence identity. The lengths of the proteins range from 50 to 200 amino acids. In the following, we perform a 4-fold cross validation procedure : for each model, the parameters are optimized on a training set made of 3/4 of the proteins and the likelihood of the remaining proteins (1/4 of the total set of proteins) is numerically evaluated. We chose this dataset because it also provides near-native decoys, which will be used in the optimization procedure (see section 7.2.3.5).

To compare the fit of our alternative models in the context provided by SC phylogenetic models, we used three phylogenetic datasets : Globin (15 vertebrate sequences of the

β-globin gene (taken from the dataset described in [Yang et al., 2000a]) with a tree topology estimated using Phylobayes 3.1c [Lartillot et al., 2009] and with the protein structure defined by the PDB file 4HHB) ; LYSIN (25 abalon sperm lysin sequences [Yang et al., 2000b], with the tree topology defined by [Yang et al., 2000b] and with the protein structure defined by the PDB file 1LYS) ; ADH (23 alcohol deshydrogenase sequences taken from Drosophila [Yang et al., 2000a], with the tree topology defined by [Yang et al., 2000a] and with the protein structure taken from the PDB file 1A4U).

7.2.3.2 Evolutionary model

The evolutionary process of a SC model between two nucleotide sequences σ and σ' can be expressed as :

$$R_{\sigma\sigma'} = Q^{mut}_{\sigma\sigma'} \cdot e^{\frac{\beta}{2}(H(s,c|\theta)-H(s',c|\theta))}, \tag{7.5}$$

where s (resp. s') is the amino acid sequence encoded by σ (resp. σ'), $H(s,c|\theta)$ is the scoring function, which depends both on the amino acid sequence s and its structure c and θ represents the parameters of the function H. For instance, if s' fits the structure less well than s, then $H(s'|\theta) - H(s|\theta)$ is positive, and the substitution is less likely to occur. The selection is modulated by the selection stringency, β. The evolutionary process is reversible, and its stationary distribution is given by [Choi et al., 2007] :

$$\varphi_{\Theta,c}(\sigma) = \frac{1}{Y_1} \cdot \Pi^{mut}_{\sigma} \cdot e^{-\beta H(s,c|\theta)}, \tag{7.6}$$

where

$$Y_1 = \sum_{\sigma'} \Pi^{mut}_{\sigma'} \cdot e^{-\beta H(s',c|\theta)}. \tag{7.7}$$

Y_1 is a normalization factor. Θ is the set of parameters of the evolutionary model (which include θ. Π^{mut}_{σ} is the stationary distribution of the mutation process, defined as :

$$\Pi^{mut}_{\sigma} = \frac{\prod_{1\leq i\leq 3n} \pi_{\sigma_i}}{1-\pi_{stop}}, \tag{7.8}$$

where σ_i represents the nucleotide at position (site) i in σ and $\pi_{stop} = \pi_{TTA}+\pi_{TAG}+\pi_{TGA}$. It is typically reasoned that a mutation leading to a stop codon is too deleterious to be accepted, and thus we do not consider mutations leading to stop codons (TAA, TAG, TGA). Note that :

$$\Pi^{mut}_{s} = \sum_{\sigma|s} \Pi^{mut}_{\sigma}, \tag{7.9}$$

where $\sigma|s$ meant that nucleotide sequence σ encodes amino acid sequence s.

We can thus obtain from this probability the stationary distribution of an amino acid sequence s :

$$\varphi_{\Theta,c}(s) = \sum_{\sigma|s} \varphi_{\Theta,c}(\sigma) = \frac{\Pi_s^{mut} e^{-\beta(H(s,c|\theta))}}{\sum_{\sigma'|s'} \Pi_{s'}^{mut} e^{-\beta(H(s',c|\theta))}} = \frac{1}{Y_2} \cdot e^{-\beta(H(s,c|\theta)+\sum_{1\leq i\leq n} \nu_{s_i}^{mut})}, \quad (7.10)$$

where we define the mutational propensity of codon c :

$$\nu_{s_i}^{mut} = \sum_{\sigma_1\sigma_2\sigma_3|s_i} -\ln \pi_{\sigma_1}^{mut} \pi_{\sigma_2}^{mut} \pi_{\sigma_3}^{mut}, \quad (7.11)$$

and

$$Y_2 = \sum_{s'\in\mathbb{S}} e^{-\beta(H(s',c|\theta)+\sum_{1\leq i\leq n} \nu_{s'_i}^{mut})}. \quad (7.12)$$

The mutation rate $Q_{\sigma\sigma'}^{mut}$ is usually derived from an i.i.d. mutation process, thus entirely characterized by a time-reversible 4x4 matrix. Here, we consider the most general time reversible mutation process, as in [Rodrigue et al., 2009]. The mutation rate between two sequences differing by more than one nucleotide is equal to 0.

7.2.3.3 Likelihood framework

Let us consider a dataset composed by p native sequence-structure pairs (i.e. proteins taken from the PDB). Then if we suppose that the proteins of the learning dataset are at the mutation-selection equilibrium [Choi et al., 2007, Choi et al., 2008], we can write the probability :

$$P(\tilde{S}|\tilde{C},\theta) = \prod_p \varphi_{\Theta,\tilde{c}^p}(\tilde{s}^p,\theta), \quad (7.13)$$

where $\tilde{s}^p$ represents the native sequence and $\tilde{c}^p$ the native structure of the p^{th} protein. This probability can be seen as a likelihood and can thus be optimized with respect to the parameters of the model. Hereafter, we will describe the method with a single protein $(\tilde{s},\tilde{c})$, although it can be easily generalized to the whole dataset. We can thus write :

$$P(\tilde{s}|\tilde{c},\theta) = \frac{1}{Y_2} \cdot e^{-\beta(H(s,c|\theta)+\sum_{1\leq i\leq n} \nu_{s_i}^{mut})}. \quad (7.14)$$

Note that we previously defined [Kleinman et al., 2006, Bonnard et al., 2009]

$$P(s|c) = \frac{e^{-H(s,c|\theta)}}{Y}, \quad (7.15)$$

but by doing so, we were not formulating the problem in terms of a mutation selection equilibrium, and for that reason, we were not correctly modeling the impact of the mutation pressure on the observed protein sequences.

The normalization factor Y_2 cannot be analytically computed. In a previous work, we addressed this problem by using Monte Carlo methods based on a Gibbs sampling algorithm [Kleinman et al., 2006]. Here, we use an approximate method instead [Bonnard et al., 2009]. Specifically, we define a pseudo-likelihood (based on the leave-one-out principle), to be optimized with respect to the set of parameters θ. Its normalization factor is analytically tractable, and we have previously shown that the method provides parameters that are virtually indistinguishable from those obtained using the MCMC framework [Bonnard et al., 2009].

To compare different models and settings, we use a cross validation test. It consists in optimizing the parameters on a dataset $D1$, and testing these parameters on another independent dataset $D2$. The likelihood obtained on $D2$ for given set of parameters and model (θ_1, M_1) can be directly compared to the likelihood obtained on $D2$, for another couple of parameters and model, (θ_2, M_2). The cross-validation score includes an intrinsic penalization of high dimensional models : if the parameters are over-fitted on D_1, they would not be adapted to describing the sequences in D_2. Here, we use a 4-fold cross validation, i.e. the dataset is partitioned so that 3/4 of the dataset is used to optimize the parameters, and the remaining quarter corresponds to the test set.

7.2.3.4 Definition of the scoring function

Many different structure-sequence scoring functions can be considered. First, one can set it equal to a pseudo-energy, depending only on how the sequence fits the native structure [Robinson et al., 2003]. Here we slightly generalize the approach, and define a scoring function consisting of two terms :

$$H(s, c|\theta) = H^{sc}(s, c) + H^0(s), \tag{7.16}$$

where $H^{sc}(s, c)$ represents how the sequence fits the native structure (depending on, but not equal to, the pseudo-energy function) and $H^0(s)$ is a structure independent selection term (e.g. some amino acids are more expensive than others). $H^0(s)$ is defined as a function of quantities analogous to chemical potentials :

$$H^0(s) = \sum_{1 \leq i \leq n} \mu_{s_i}. \tag{7.17}$$

Both $H^{sc}(s, c)$ and $H^0(s)$ depend on their own set of parameters which are included in θ. However, for notation simplicity, we will hereafter omit the dependence to the parameters for these two functions.

As for the pseudo-energy function $E(s, \tilde{c})$ (on which $H^{sc}(s, c)$ depends), we use knowledge-based statistical potentials, based on a coarse grained description of the structure. This choice is motivated by at least two reasons. First, the entire SC framework entails repeated evaluations of $H(s, c|\theta)$, and therefore, the computations have to be as inexpensive as possible. Second, there is an indirect advantage in using a coarse-grained representation, as it makes the model implicitly more tolerant to slight changes of the conformation along the evolutionary lineages. Specifically, we used the same statistical potential as in [Kleinman et al., 2006, Bonnard et al., 2009] :

$$E(s, \tilde{c}) = \sum_{1 \leq i < j \leq n} \Delta_{ij}^{\tilde{c}} \varepsilon_{s_i s_j} + \sum_{1 \leq i \leq n} \sum_{1 \leq d \leq D} B_{i,d}^{\tilde{c}} \alpha_{s_i}^{d}, \tag{7.18}$$

where $\Delta^{\tilde{c}}$ is the contact map of the native structure $\tilde{c}$, defined so that $\Delta_{ij}^{\tilde{c}} = 1$ if the sites i and j are distant by less than a contact threshold. $B^{\tilde{c}}$ is the solvent accessibility map with $B_{i,d}^{\tilde{c}} = 1$ if the site i belongs to the solvent accessibility class d ($d = 1..D$). These terms only depend on the structure $\tilde{c}$, ε_{ab} representing the contact energy between the two amino acids a and b and α_a^d the solvent accessibility energy of the amino acid a in the solvent accessibility class d. These ε_{ab} and α_a^d terms are parameters of the statistical potential that we will optimize.

Equation (7.18) can be generalized [Kleinman et al., Submitted] to more than one contact map, which thus provides a way of representing a distance-dependent pairwise energy function : for each possible contact map m, $\Delta_{ij}^{\tilde{c},m} = 1$ if i and j are distant by more than a threshold t_{min}^m and less than another, t_{max}^m, so that $t_{min}^m = t_{max}^{m-1}$ and $t_{max}^m = t_{min}^{m+1}$. Thus, the three-dimensional space can be partitioned into different distance maps [Kleinman et al., Submitted], and to each distance map $\Delta^{\tilde{c},m}$ corresponds a distance potential ε_{ab}^m.

Each amino acid in a protein structure is only represented by its center of mass. The contact map is then determined by considering that two sites are in contact if their center of mass are distant by less than a contact threshold. We choose two different representations, using one (threshold = 6.5 Å) or three (thresholds = 4.5, 6.5 and 8.5 Å) contact maps. For the accessibility map, we used Naccess [Hubbard and Thornton, 1993] to determine the solvent accessibility area of each site, and then partitioned the space of solvent accessibility areas into 14 classes [Kleinman et al., 2006].

7.2.3.5 Accounting for preference against decoys

Optimizing the parameters so that the native protein, i.e. sequence-structure pair $(\tilde{s}, \tilde{c})$, has the lowest energy among all possible couples $(s, \tilde{c})$ corresponds to finding the

parameters that maximize the stability of the protein. Here, rather than the stability, we want instead to maximize the specificity of the protein, i.e that the native protein has lower energy than other (s, c) couples.

The problem can be formalized as follows. At thermodynamic equilibrium, the probability distribution of a given sequence to be in a three-dimensional structure is given by the Boltzmann distribution :

$$f(c|s) = \frac{\exp\left(-\beta E(s,c)\right)}{Z_s}, \tag{7.19}$$

where

$$Z_s = \sum_{c' \in \mathbb{C}} \exp\left(-\beta E(s,c')\right), \tag{7.20}$$

where $\beta = 1/kT$ is the inverse temperature (not to be confused with the selection stringency in eq. (7.4)). For any possible structure c, $f(\hat{c}|s,\theta) \leq f(c|s,\theta)$, so that $\hat{c}$ is the *ground state* of s.

Conversely, finding a sequence s that has $\hat{c}$ as its ground state, and such that the specificity is maximum, amounts to maximizing $f(\hat{c}|s,\theta)$, or equivalently, to minimizing $-\ln f(\hat{c}|s)$ [Seno et al., 1996]. Therefore, we can define the scoring function $H^{sc}(\tilde{s},\tilde{c})$:

$$H^{sc}(\tilde{s},\tilde{c}) = -\ln f(\tilde{c}|\tilde{s}) = E(s,\hat{c}) + \ln Z_s. \tag{7.21}$$

The normalization factor $\ln Z_s$ cannot be analytically computed, and we therefore have to resort to approximations.

Expanding $\ln Z_s$ in powers of β, we obtain [Deutsch and Kurowski, 1996] :

$$\begin{aligned} \ln Z_s(\beta) &\simeq \ln Z_s(0) + \beta \sum_{c' \in \mathbb{C}} E(s,c') \cdot f(c'|s,\theta) \\ &\simeq \ln Z_s(0) + \beta \langle E(s,c') \rangle_{c' \in \mathbb{C}}. \end{aligned} \tag{7.22}$$

Thus, up to an additive constant which we will ignore in the following, the term $\ln Z_s$ can be approximated by the average energy of the sequence when folded into a set of alternative structures, which are themselves drawn from the Boltzmann distribution given by eq. (7.19). Note that this distribution depends on the current sequence. In addition, it is difficult to sample from this distribution (as this would require an efficient search through the conformational space of a given protein, which is known to be virtually intractable). An alternative is to replace this distribution by an externally defined set of decoys. Hence, solving the problem amounts to finding an adequate set of decoys. In the following, we will compare alternative decoy sets, using empirical tests based on cross-validation and Bayes factors.

Gathering equations (7.16), (7.17) and (7.22), the overall scoring function reads as :

$$H(\tilde{s},\tilde{c}|\theta) = \underbrace{E(\tilde{s},\tilde{c}) - \langle E(s,c')\rangle_{c'}}_{H^{SC}(s,c)} + \sum_{1\leq i\leq n} \mu_{s_i}. \tag{7.23}$$

The random energy model

The simplest approximation is the one we used in our previous works [Kleinman et al., 2006,Bonnard et al., 2009,Kleinman et al., Submitted]. It is inspired by the random energy model (REM) [Finkelstein, 1997, Shakhnovich and Gutin, 1993] :

$$\langle E(s,c')\rangle_{c'} = \sum_{1\leq i\leq n} \lambda_{s_i}, \tag{7.24}$$

where λ_a are parameters depending only on the amino acid a. The REM model assume that the structure space can be approximated by a random mixture of amino acids and that the average energy depends mainly on the composition of the sequence. The scoring function of the SC^{REM} model is then :

$$H^{REM}(\tilde{s},\tilde{c}|\theta) = E(\tilde{s},\tilde{c}) + \sum_{1\leq i\leq n} \lambda_{s_i} + \sum_{1\leq i\leq n} \mu_{s_i}. \tag{7.25}$$

Note that, when using the REM approximation, the two terms λ_a and μ_a cannot be separately identified. Nor can they be identified from solvent accessibility energies in the optimization procedure. As in our previous work, we therefore consider them as implicitly embedded in the accessibility parameters [Bonnard et al., 2009] ($\alpha'^d_a = f(\alpha^d_a, \lambda_a, \mu_a)$).

Inclusion of explicit decoys

Alternatively, the expectation $\langle E(s,c')\rangle_{c'}$ can be approximated by an explicit average over a finite set of decoys. The linearity of the equation allows some computational shortcuts :

$$\begin{aligned}
H^{dec}(\tilde{s},\tilde{c}) &= E(\tilde{s},\tilde{c}) - \langle E(s,c')\rangle_{c'} + \sum_{1\leq i\leq n} \mu_{s_i} \\
&= \sum_{1\leq i<j\leq n} \Delta^{\tilde{c}}_{ij}\varepsilon_{s_i s_j} + \sum_{1\leq i\leq n}\sum_{1\leq d\leq D} B^{\tilde{c}}_{i,d}\alpha^d_{s_i} - \sum_{1\leq i<j\leq n} \langle\Delta^{c'}_{ij}\rangle_{c'}\varepsilon_{s_i s_j} \\
&\quad - \sum_{1\leq i\leq n}\sum_{1\leq d\leq D} \langle B^{c'}_{i,d}\rangle_{c'}\alpha^d_{s_i} + \sum_{1\leq i\leq n} \mu_{s_i} \\
&= \sum_{1\leq i<j\leq n} \left(\Delta^{\tilde{c}}_{ij} - \langle\Delta^{c'}_{ij}\rangle_{c'}\right)\varepsilon_{s_i s_j} + \sum_{1\leq i\leq n}\sum_{1\leq d\leq D} \left(B^{\tilde{c}}_{i,d} - \langle B^{c'}_{i,d}\rangle_{c'}\right)\alpha^d_{s_i} + \sum_{1\leq i\leq n} \mu_{s_i}.
\end{aligned} \tag{7.26}$$

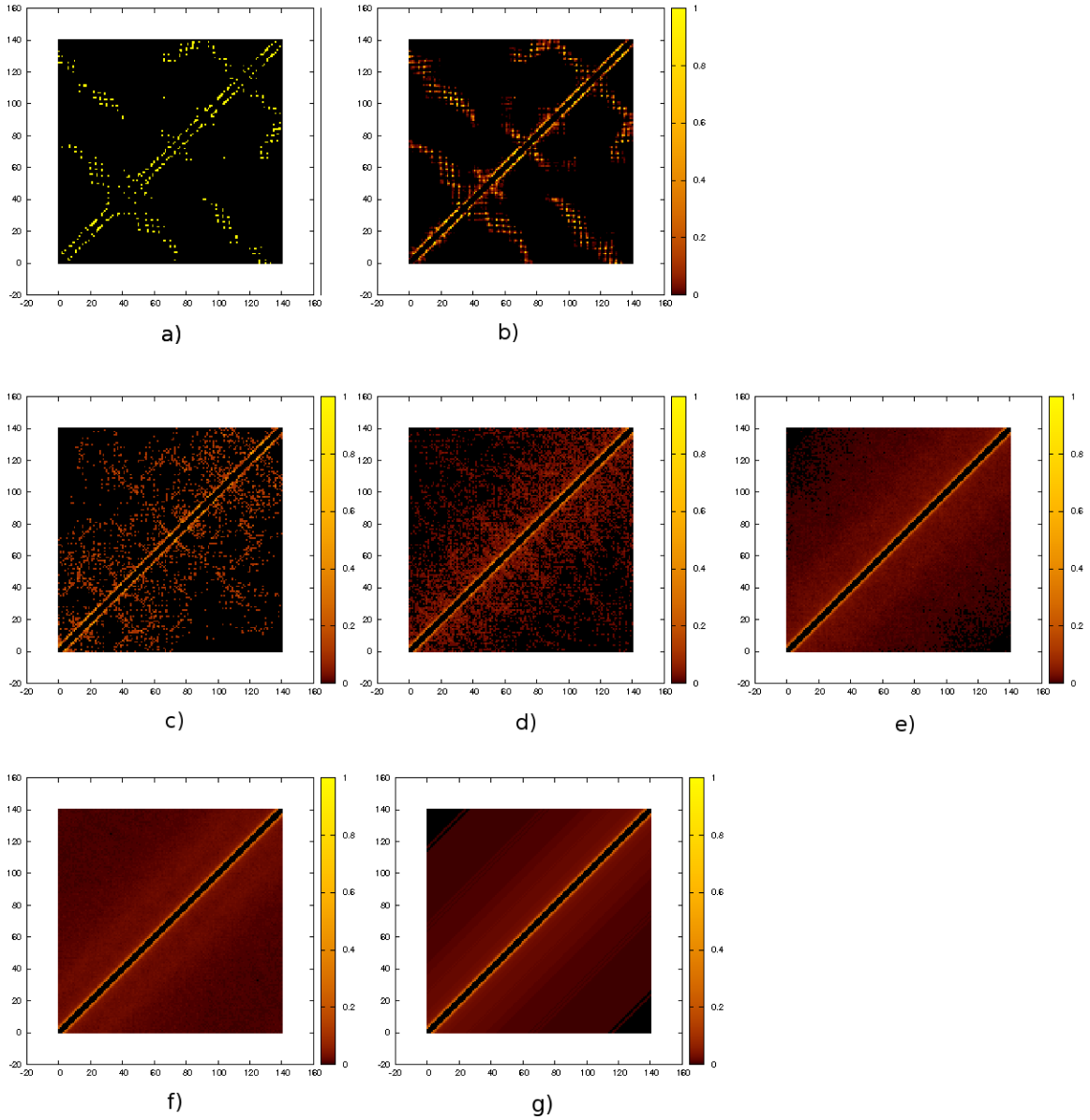


FIGURE 7.1 – The average contact maps obtained for the B-chain of hemoglobin (from Dasyatis akajei, a cartilaginous fish) : a) native structure, b) decoys from DS^{hrd}, c) from DS_{10}^{gap}, d) from DS_{50}^{gap}, e) from DS_{1000}^{gap}, f) from DS^{thr}, g) from DS^{ave}.

Hereafter, we will call $\langle \Delta_{ij}^{c'} \rangle_{c'}$ the *average decoy contact map* and $\Delta_{ij}^{\tilde{c}} - \langle \Delta_{ij}^{c'} \rangle_{c'}$ the *effective contact map*. This latter term can be understood as a contrast between the structural description of the native conformation, and the average structural description over the decoy set. This definition of the selection function will be used with different explicit decoys, leading to the following models : a previously defined high resolution decoy dataset (SC^{hrd}), and decoys obtained by threading (SC^{gap}, SC^{thr} and SC^{ave}).

First, we will test a *high resolution decoy* dataset, made of misfolded, near-native structures. We took the *high resolution decoy* dataset defined by [Rajgaria et al., 2008]. This dataset DS^{hrd} was constructed using the following procedure : first, for each protein, the secondary structure and the hydrophobic core were determined ; second, the torsion angles of the defined structure were modified using DYANA [Günter et al., 1997] ; and third near native structures were selected (mean rmsd =2.02 Å) [Rajgaria et al., 2008]. For each protein, the dataset contains about 500 decoys.

The second family of datasets that we will test is made of decoys obtained by threading, i.e. the native sequence is artificially folded into real structures, taken from the PDB. These decoys are supposed to be a representative sample of the whole set of possible structures, and not only the near-native decoys (which are represented by the dataset DS^{hrd}). We used two different threading methods. The first is a gapped threading [Finkelstein, 1997], on the dataset itself, allowing for chimeric proteins. To do this, for each protein p (with length n), we randomly choose a site k in a protein $p' \neq p$ of the dataset and take the n following sites to construct the contact map. If there are not enough sites in protein p' between site k and the C terminus of the protein to fulfill the contact map, we take the remaining sites from the next protein $(p' + 1)$ in the database. To test how many structures are needed in the decoy set, we used three datasets composed of 10, 50 and 1000 decoys per protein, which we call DS_{10}^{gap}, DS_{50}^{gap}, and DS_{1000}^{gap}. The second method is the gapless threading [Finkelstein, 1997] on an independent dataset (made of more than 3000 proteins) leading to the dataset DS^{thr}. The number of decoys per protein varies between 1804 and 3363, as the length of the threaded sequences is sometimes larger than the length of some proteins of the threading dataset.

To understand how the decoy contact map may depend on the set of decoys, we represented for one protein (with a threshold of 6.5 Å), the contact map of the native structure and the averaged contact maps using the different methods introduced above. As shown in fig. 7.1, the decoys of the dataset DS^{hrd} (fig. 7.1.b) are near-native. They show the same secondary and three dimensional structures as the native protein (fig. 7.1.a), albeit with a greater variability. In contrast, the decoy datasets obtained by threading, DS^{gap} (fig. 7.1.c-e) and DS^{thr} (fig. 7.1.f), seems to be much more spread out in structure

space, resulting in a flatter average decoy contact map. We observed that the datasets DS_{10}^{gap} and DS_{50}^{gap} present a lot of variability in their contact maps (data not shown), because of the small number of decoys for each protein.

The last contact map (fig. 7.1.g), DS^{ave}, can be seen as the limit of the contact maps constructed by gapped and gapless threading with an increasing number of decoys. Fig. 7.1 illustrates the fact that if we include an unlimited number of decoys obtained by threading in the dataset DS^{thr}, the resulting contact map will be such that the average contact frequency between i and k will only depend on $t = |k - i|$. This DS^{ave} contact map is thus entirely characterized by a single-row vector χ_t, containing, for each t, the averaged contact frequency between sites separated by t positions along the sequence. A similar observation was made by [Rossi et al., 2001].

Variational method

Based on the translation invariance just observed (i.e. χ_t is shared by every sites), a new variational method can be constructed. This corresponds to the variational decoy contact map, in which the χ_t are not calculated on the dataset, but are optimized along with the parameters of the statistical potential. The potential defined in this way is called the variational potential. The optimized χ_t will correspond to a 'maximum likelihood virtual decoy set' under the constraint of translational invariance.

By analogy, we can construct a variational decoy set for the solvent accessibility term : each site has the same propensity, in the decoys, to be in a given solvent accessibility class, which is equal to the average fraction of sites observed in this class. In a variational context, this fraction can be represented by a parameter ζ_d, and ζ is a single-row vector.

As χ_t represents an average decoy contact map, we have the following constraint :

$$0 \leq \chi_t \leq 1, \ \forall t = 3..T, \tag{7.27}$$

where t vary from 3 (because we do not consider contact for sites distant by less that 3 amino acids) to T, T being a predefined maximum, $T = 50$.

Altogether, we propose the following scoring function :

$$H^{var}(s, c|\theta) = \sum_{1 \leq i < j \leq n} (\Delta_{ij}^{\tilde{c}} - \chi_{|j-i|})\varepsilon_{s_i s_j} + \sum_{1 \leq i \leq n} \sum_{1 \leq d \leq D} (B_{i,d}^{\tilde{c}} - \zeta_d)\alpha_{s_i}^d - \sum_{1 \leq i \leq n} \eta_{s_i}. \tag{7.28}$$

Phylogenetic comparison

The phylogenetic SC model used here has been described previously [Rodrigue et al., 2006, Rodrigue et al., 2009]. In the present case, we just plug-in the alternative statistical potentials we have estimated into this framework, and measure the fit by Bayes factor

computations. The Bayes factor between two models M_1 and M_2 is defined as the ratio of the marginal likelihoods of the two models under comparison :

$$B(M_0, M_1) = \frac{P(A|M_1)}{P(A|M_0)}, \tag{7.29}$$

where

$$P(A|M_1) = \int_\theta p(A|\theta, M_1)p(\theta|M_1)d\theta \tag{7.30}$$

is the marginal likelihood of model M_1 (the parameters are integrated over the entire prior distribution), and A represents the data (alignment of nucleotide sequences). A Bayes factor higher than 1 is considered as an evidence in favor of M_1. As described in [Rodrigue et al., 2006], Bayes factors can be estimated by thermodynamic integration (*path sampling*).

Note that the potentials obtained here already include a mutational component (ν^{mut}, eq. (7.10) and (7.11), which is also accounted for by the phylogenetic SC model. To avoid counting this component twice, we retrospectively subtract a term ν^{mut} defined by eq. (7.11), based on an estimate of the empirical frequencies of the nucleotides at the third position for four-fold degenerate sites in the PDB.

7.2.4 Results

7.2.4.1 Optimization of the parameters

For each dataset, we perform different optimization runs, with randomly chosen initial parameter values, and we checked that the optimization leads to the same final values of the parameters and that they are biologically plausible. If we compare the parameters obtained under the random energy model model (SC^{REM} fig. 7.1.a) and using the high-resolution decoy model (SC^{hrd}), we observe a weak correlation ($R^2 = 0.67659$) for the contact potential parameters. The correlation is more evident ($R^2 = 0.96316$) between the parameters obtained using the random energy model and the gapless threading method (SC^{thr}, fig. 7.1.b). The higher correlation between SC^{REM} and SC^{thr} than between SC^{REM} and SC^{hrd} probably indicates that the random energy approximation and the threading method both attempt to address the specificity problem at a similar level of coarse graining. In contrast, the decoys of Rajgaria et al [Rajgaria et al., 2008] clearly represent a more fine grained representation of the competition between native and non-native structures (fig. 7.1), which translates into less similar parameters upon optimization (fig. 7.2).

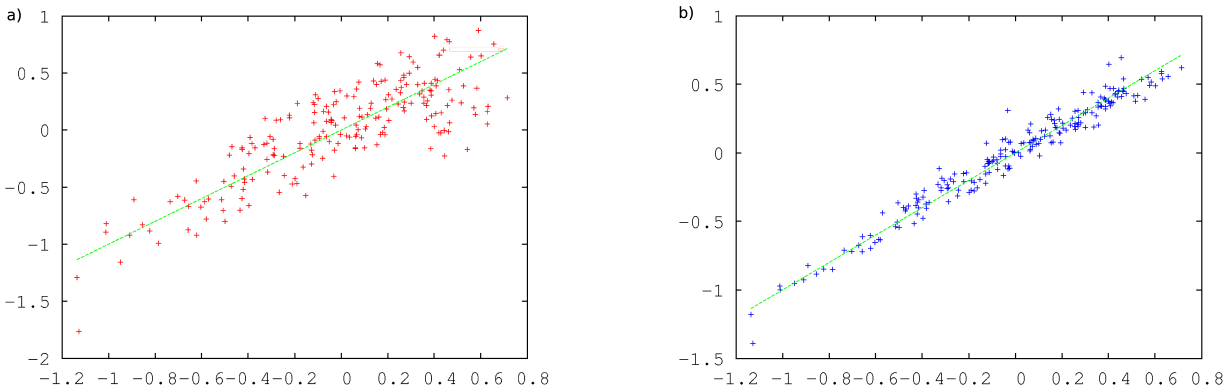


FIGURE 7.2 – XY-plot of the contact potentials. X-axis : potential obtained using the random energy model (DS^{REM}) Y-axis : potential obtained from DS^{hrd} (in red) and the dataset DS^{thr} (in blue).

In the case of the variational method, the average decoy contact map belongs to the parameters that are optimized by maximum likelihood. In fig. 7.3, we display the resulting values, along with the observed values of DS^{ave}. The X-axis represents the distance in the sequence between two sites, and the Y-axis represents the percentage of contacts observed or inferred in the decoys. We observe that the two distributions are quite different. Most strikingly, the variational method leads to a very low proportion of contact frequencies for $|j-i| \geq 10$, apart from a few entries where the value of the SC^{var} decoy contact map is higher than for the SC^{ave} model (between 1.5 and 3 times the proportion observed for the SC^{ave} model). In comparison, the proportion of observed contacts in the decoys for the SC^{ave} model slightly decreases for $|j-i| \geq 9$. The threshold of 16 used by Rossi et al [Rossi et al., 2001] may be sufficient, because the proportion of contacts observed between sites separated by more than 6 sites, is less than 5%, suggesting that as optimization of the variational potential using this threshold may be sufficient. Optimizing the potential using a threshold of 16 would also allow us to assess whether the peaks observed in fig. 7.3 are important for the fit of the SC^{var} model.

7.2.4.2 Cross validation

We systematically tested the alternative models by cross validation, optimizing the parameters on a dataset, and testing the fit on an independent dataset (see methods). To make the reading easier, table 7.1 presents the cross validation score $-\ln P(\tilde{s}|\tilde{c},\theta)$

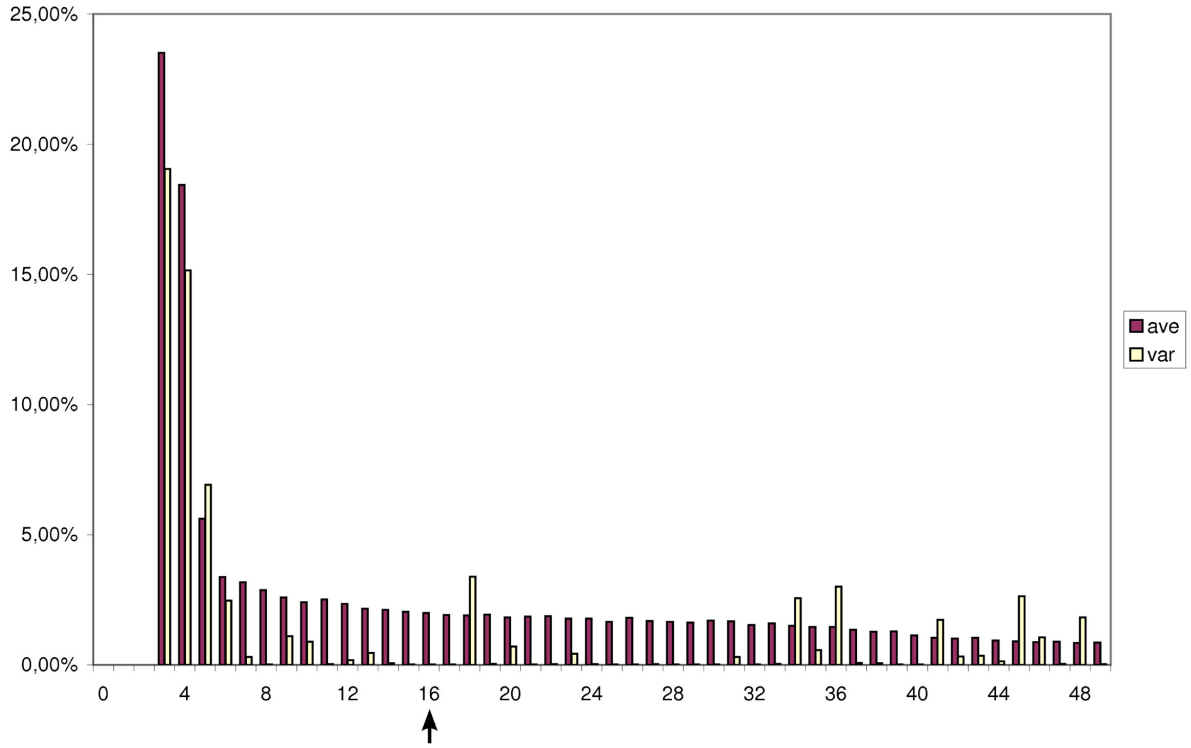


FIGURE 7.3 – Comparison of the averaged contact map (SC^{ave}) and the variational contact map (SC^{var}) models, as a function of the distance in the sequence. The arrow indicates the threshold used by [Rossi et al., 2001].

per site (better models having lower scores). Note that because of the variability of the structures of the decoy sets constructed by gapped and gapless threading, we performed four optimization runs for each dataset (DS_{10}^{gap},DS_{50}^{gap}, DS_{1000}^{gap} and DS^{thr}), computed the cross validation score (for each test set, we perform five cross-validation runs for each optimization run) and averaged the cross validation score over the 20 runs.

The most important conclusions that can be drawn from this general cross-validation analysis are as follows :

- First, the DS^{hrd} dataset, composed of high resolution, near-native decoys, does not result in an improvement of the fit, compared to the random energy approximation (DS^{REM}). This result may seem surprising at first, because including near-native decoys is supposed to be an important feature of any method aiming at accounting for specificity. On the other hand, some explanations can be proposed. In particular, our statistical potential is probably too coarse-grained to efficiently capture the small differences (see fig. 7.1) between the native and the decoys provided by Rajgaria et al. These

	M = 1	M = 3
DS^{REM}	-2.66100	-2.61164
DS^{hrd}	-2.69448	-2,65161
DS^{gap}_{10}	-2.67523	-2,62501
DS^{gap}_{50}	-2,66293	-2.61164
DS^{gap}_{1000}	-2,65956	-2.60800
DS^{thr}	-2,65977	-2.60838
DS^{ave}	-2.65958	-2.60795
DS^{var}	-2.65829	-2.59890

TABLE 7.1 – Cross-validation : For each type of contact map definition (one or three contact map) we present the average score per site.

high-resolution decoys are probably more suitable for the problem of modeling side-chain packing. The poor fit obtained in the present case clearly indicates that they are unsuited for our coarse-grained model.

- Second, and as expected, the fit of the decoy sets obtained by gapped threading increases with the number of structures included in the decoy set. For low numbers (10 to 50), the fit is worse than that of the random energy approximation, while it is better for 1000 decoys. The score stabilizes for a set of 500 to 1000 decoys per protein (fig. 7.4).

- Third, the scores obtained for the three methods based on threading (*thr*, gap_{1000} and *ave*) are very similar, which indicates that the specific method for performing the threading is rather unimportant. One can expect that the dataset DS^{thr} is a sufficient sample of the whole set of plausible decoys. Indeed, it is obtained by the gapless threading of the sequence of interest into real protein structures, even if they are not near the native structure of the threaded sequence. The cross-validation score obtained is better than for the dataset DS^{REM}, but the improvement is rather modest, compared to that observed between a simple contact potential and a solvent accessibility and contact potential only (whose cross validation score is -2.69648).

- And finally, the fit of the variational method is slightly better than that of the threading methods. A better fit was expected, given that the variational method simply generalizes the functional form of a typical average of the pairwise and the accessibility

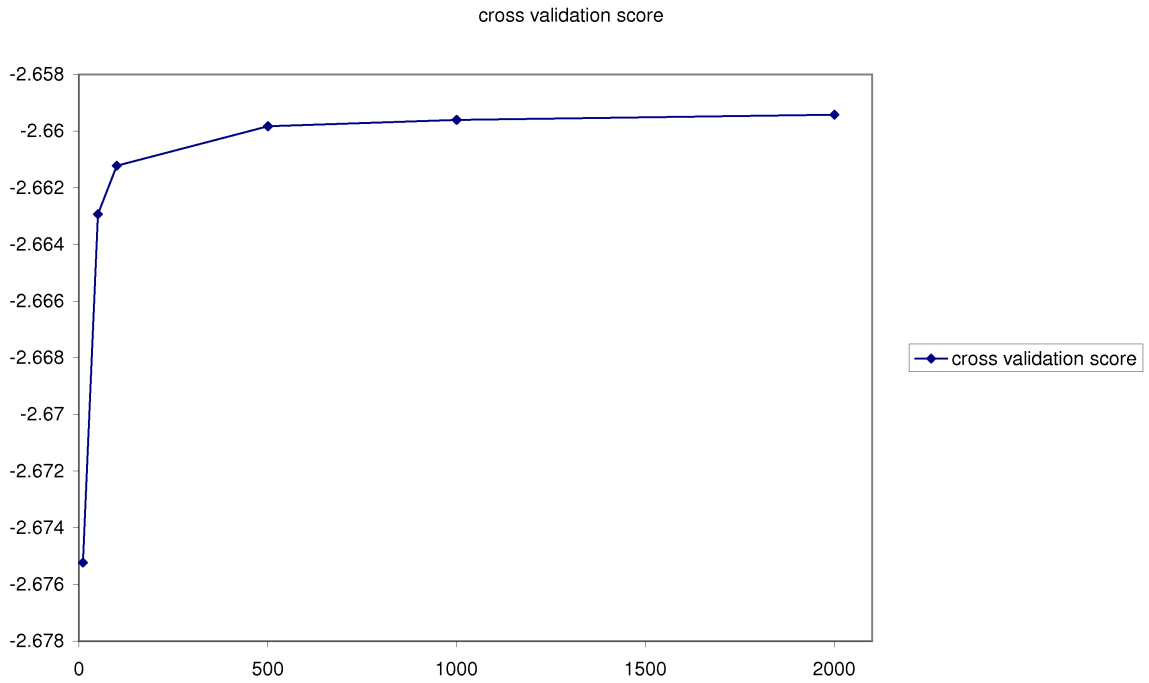


FIGURE 7.4 – Evolution of the cross validation score per site as a function of the number of decoys included for each protein (one contact map).

potentials on a threading set. Conversely, and importantly, the fact that the improvement is very small indicates that a simple implementation of the threading approach, based on a random sampling of natural proteins, as we have done here, seems to already yield the best results that this method can achieve.

7.2.4.3 Phylogenetic comparison

The alternative statistical potentials obtained above were all included in a SC phylogenetic model, implemented in a general Bayesian Monte Carlo framework [Rodrigue et al., 2005, Rodrigue et al., 2006, Rodrigue et al., 2009]. Specifically, the substitution process defined by :

$$R_{\sigma\sigma'} = Q^{mut}_{\sigma\sigma'} \cdot e^{\frac{\beta}{2}(H(s,c|\theta)-H(s',c|\theta))}, \tag{7.31}$$

and using each of the statistical potentials previously optimized on PDB in the score function $H(s, c|\theta)$, is assumed to operate along the lineages of a phylogeny, whose topology is fixed a priori.

The parameter β in eq (7.31) can be seen as a parameter tuning the overall stringency of the structure-dependent selection, such as modeled by our statistical potential. When

		SC^{REM}	SC^{thr}	SC^{var}
Adh	Bayes factor	[142.16 : 142.258]	[144.287 : 145.982]	[146.431 : 146.086]
	β	[0.88]	[0.91 : 0.92]	[0.89 : 0.9]
bGlobin	Bayes factor	[123.095 : 126.145]	[130.152 : 130.81]	[126.462 : 127.737]
	β	[0.95]	[1.02]	[0.95 : 0.99]
Lys	Bayes factor	[52.965 : 53.148]	[55.927 : 56.386]	[51.843 : 52.808]
	β	[0.69 : 0.71]	[0.72 : 0.73]	[0.7 : 0.71]

Table 7.2 – Maximal natural logarithm of the Bayes factors, and the corresponding optimal values of β, based on a structural description using one contact map.

		SC^{REM}	SC^{thr}	SC^{var}
Adh	Bayes factor	[159.945 : 160.761]	[159.603 : 161.039]	[155.786 : 156.249]
	β	[0.83 : 0.84]	[0.85 : 0.86]	[0.83 : 0.85]
bGlobin	Bayes factor	[119.063 : 120.35]	[126.69 : 127.403]	[118.542 : 118.549]
	β	[0.85]	[0.87 : 0.89]	[0.83 : 0.85]
Lys	Bayes factor	[67.3635 : 67.6714]	[73.183 : 73.975]	[75.445 : 76.636]
	β	[0.69 : 0.72]	[0.75 : 0.79]	[0.78 : 0.81]

Table 7.3 – Maximal natural logarithm of the Bayes factors, and the corresponding optimal values of β, based on a structural description using three contact maps.

$\beta = 0$ the model reduces to the pure mutation model PM_0. A value of β around 1 is expected, given the way the parameters were optimized. However, β can vary between proteins. In addition, applying the path sampling method along the entire interval $\beta \in [0 : 2]$ allows us to numerically evaluate the fit of the model as a function of β [Rodrigue et al., 2006].

As an illustration, fig. 7.5 shows the evolution of the natural logarithm of the Bayes factors (against the pure mutational model) for the SC^{REM}, SC^{var}, and SC^{thr} models, based on the numerical evaluation by path sampling.

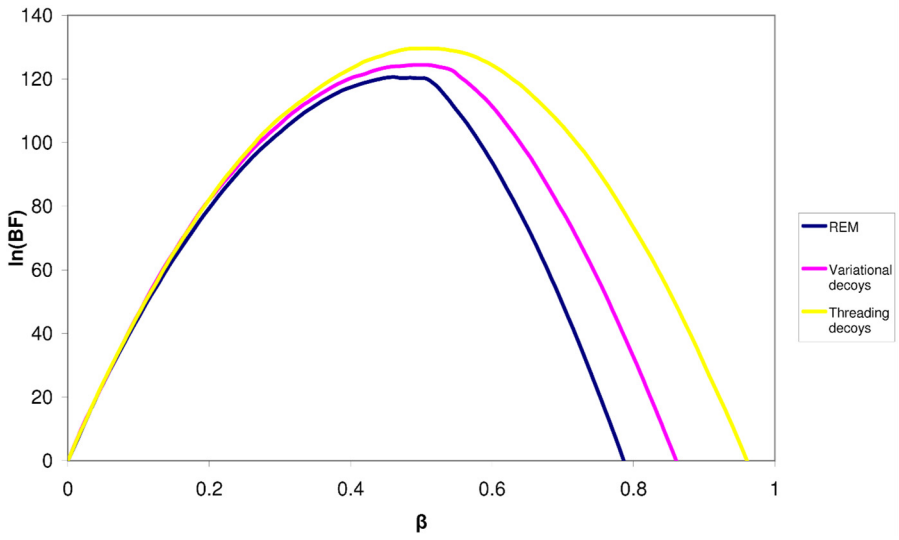


FIGURE 7.5 – Evolution of the logarithm of the Bayes factors as a function of β for the dataset BGLOBIN, and a structural description using one contact map, for the SC^{REM} model (dark blue), the SC^{var} model (pink) and the SC^{thr} model (yellow).

In table 7.2 and 7.3 we report the maximal natural logarithm of the Bayes factor with its corresponding selection intensity parameter β. As a general rule, and both for the contact (table 7.2) and distance (table 7.3) pairwise potentials, the fit of at least one decoy-based models is higher than that of the random energy model. Whether the variational method is better than the threading method is much less evident, as no trend is apparent in tables 7.2 and 7.3.

The optimal values of β are always correlated with the fit of the models, with better models resulting in a value of β closer to 1, and less fit models a lower value of β. For instance, the SC^{REM} model seems to be less adapted to the LYS dataset ($\beta \in [0.69 : 0.72]$) while the variational model seems to be more adapted to this dataset ($\beta \in [0.78 : 0.81]$).

7.2.5 Discussion

7.2.5.1 Inclusion of explicit decoys

In this article, we present a general framework including an explicit modeling of the conformational specificity requirement in structurally constrained models of molecular evolution. We compare different decoy datasets by cross validation and incorporate some of them into a SC model.

The inclusion of decoys is arguably one of the most important features that should be incorporated in a protein-design or a structurally constrained evolutionary framework. The cross validation tests we performed here indicate that accounting for specificity using decoys indeed results in a better fit. On the other hand, the improvements we were able to achieve are somewhat modest. The main reason is probably that the decoy sets that we used here are not adapted to the pseudo-energy function and to the coarse-grained structural representation assumed by this energy function.

On one hand, the decoys of Rajgaria et al (DS^{hrd}, fig. 7.1.b) are probably too close to the native structure to be informative in the coarse grained context of our statistical potential. Nevertheless, it would be difficult to test any other existing decoy structures datasets, because most of these datasets contain a lot of decoys, but for few different native structures, and thus cannot be used in our learning method, which demands large datasets both in terms of decoys and in terms of number of distinct natural proteins. Moreover, loop inversion, or variation of the native structure by molecular dynamics gives structures that would be too near-native to be informative in our coarse-grained model, making these types of decoys unsuitable for the present context.

On the other hand, decoys constructed by threading, although they seem to yield better results, appear to be too unrelated to the native structures. In the light of these observations, classical methods to construct decoy structures datasets do not seem to be adapted to the present context, which raises the question whether we could identify methods for constructing decoys that would be better tailored to the present needs.

In this direction, interesting clues are provided by the theory itself, if we come back to the approximation of $\ln Z_s$:

$$\begin{aligned} \ln Z_s &\simeq \langle E(s,c') \rangle_{c'} \\ &\simeq \sum_{c'} p(c'|s) E(s,c'). \end{aligned} \tag{7.32}$$

According to this equation, the decoy set is supposed to be a representative sample of the distribution probability $p(c'|s)$. It is clear that the proposed decoy dataset do not offer

a very good approximation of this Gibbs ensemble. Equation (7.32) also suggests that the decoy datasets should vary according to the proposed potential, and according to the current sequence being evaluated. Of course, we cannot efficiently sample such decoys from the whole set of structures (except for small lattice models), but this suggest that we could create new decoy datasets spanning a wide range of variations, from very close to to very far from the native structure (e.g. containing DS^{hrd} and DS^{thr}), and then re-weight this set according to some distance between each decoy and the current native structure. However, we have to keep in mind that the number of decoys in the dataset would be a limiting factor for this method.

7.2.5.2 The variational method

Rossi et al. [Rossi et al., 2001] assume that the decoy contact maps are invariant by translation (i.e. the frequency of contacts between amino acids i and $i+k$ only depends on k). This property is naturally implied by the threading procedure. The variational method introduced here proceeds from a similar argument, although reformulated in a likelihood framework. In practice, it appears to be very similar to threaded decoys. In particular, the fit of the variational method, as measured by cross-validation, is not fundamentally different from that of the threading method.

Nevertheless, the variational method which we have introduced here is interesting for several reasons. First, seen as a theoretical limit to the threading procedure, it provides an upper bound over threaded sets of decoys meant for a general use (i.e. not specifically adapted to the native structure being considered). In particular, this implies that, in our context, substantial progress cannot be achieved by just choosing a better set of proteins from which to derive the set of decoys by threading.

Second, the variational method can be generalized to other settings. For instance, one could devise a semi-variational method in which the impact of the decoys on the scoring function could be modulated using additional parameters. For instance, a parameter $0 \leq \tau \leq 1$ could be introduced, such that the effective contact map becomes $\Delta_{ij}^{\tilde{c}} - \left(\langle \Delta_{ij}^{c'} \rangle_{c'}\right)^{\tau}$. Note that, because that $0 \leq \langle \Delta_{ij}^{c'} \rangle_{c'} \leq 1$, a value of τ close to 0 means that the influence of the decoys is emphasized, while large values of τ would dampen out the effect of the decoys on the computations.

Many other possibilities could be explored. The general idea is always the same : one first describes the set of decoys in its broad lines, up to a few unknown parameters, which are then optimized by maximum likelihood, along with the other parameters of the model. The first step can be inspired from arguments ranging from the statistical physics

of random coils, to normal mode analysis or to the empirical observation of the thermal fluctuations of proteins in NMR.

Of course, the statistical potential also needs a more complex formulation than that proposed here, and slight changes in the description of the contact maps may lead to different optimal decoy sets. For example, the improvement obtained by accounting for specificity is better for the distance potential than for the single-map contact description. Thus far, improving the statistical potential itself [Kleinman et al., Submitted] has resulted in a much greater pay off than trying to account for decoys, in term of cross validation. However, SC models using potentials are still outperformed by models using a simple synonymous/non synonymous rate ratio. Thus, as the improvement offered by the two orientations are conceptually independent from each other, new forms of potentials, as well as decoys descriptions, will have to be developed simultaneously.

7.3 Conclusion

Au premier abord, les résultats de validation croisée semblent indiquer que les jeux de données de structures leurres ne sont pas adaptés aux potentiels statistiques que nous avons étudié ici. Cependant, à l'aide d'un partitionnement différent de l'espace (trois matrices de contact au lieu d'une seule), les potentiels statistiques optimisés en prenant en compte les structures leurres (construites par *threading* ou les structures virtuelles) semblent mieux adaptées au modèle d'évolution soumis à des contraintes structurales. Ceci est intéressant car un potentiel de distance discrétisée semble mieux refléter les relations tri-dimensionnelles liant les différentes positions de la protéine.

Il serait également intéressant de comparer les modèles présentés ici au modèle présenté dans [Bastolla et al., 2006]. S'il est impossible de comparer directement les deux modèles d'évolution dans leurs contextes respectifs, il est cependant possible d'effectuer la manipulation des potentiels statistiques présentée par [Bastolla et al., 2006] (cf. section 2.2.6.4), puis de les intégrer dans le modèle d'évolution utilisé ici. Par une évaluation numérique du facteur de Bayes, il sera alors possible de savoir si l'approche en question est plus performante. A première vue, il est assez probable que les performances soient équivalentes à celles observées ici pour les structures leurres obtenues par *threading*. Le facteur déterminant reste probablement le choix des structures leurres.

Il est intéressant également de comparer les deux améliorations principales observées sur le modèle. L'intégration de nouveaux termes (torsion, B-factor...) de potentiel a permis d'obtenir d'excellents résultats de validation croisée, bien meilleurs que ceux obtenus en intégrant des structures leurres dans le processus d'optimisation. A noter qu'en outre le

modèle proposé par [Kleinman et al., Submitted] n'inclus pas le terme correctif présenté au chapitre 6, et il est possible que dans un tel contexte, le nouveau potentiel présenté par C.L. Kleinman soit plus performant qu'il ne se semble actuellement.

Il serait intéressant d'optimiser le potentiel complexe [Kleinman et al., Submitted] en utilisant la méthode proposée dans ce chapitre. Cependant, si l'amélioration proposée par l'approximation *leave-one-out* présentée au chapitre 5 permet d'améliorer la complexité des algorithmes utilisés, l'intégration des structures leurres (et plus encore celle des structures virtuelles) devient relativement gourmand en terme de temps de calcul, à cause de l'augmentation drastique du nombre de contacts.

Troisième partie

Bilan de l'approche

Chapitre 8

Perspectives

Dans cette thèse, j'ai présenté une méthode entièrement probabiliste d'optimisation de potentiels statistiques pour des modèles d'évolution de protéines soumises à des contraintes structurales. Dans le cadre probabiliste que nous avons défini (chapitre 4), je me suis orientée vers des améliorations techniques de la méthode d'optimisation, en définissant une méthode d'optimisation plus rapide, basée sur une pseudo-vraisemblance (chapitre 5), puis en essayant de prendre en compte des structures leurres dans la procédure d'optimisation et dans le modèle d'évolution. D'un autre côté, C.L. Kleinman s'est orientée vers la définition de nouvelles formes de potentiels. Si les résultats actuellement obtenus ne permettent pas de surpasser les meilleurs modèles phénoménologiques existants, ils sont encourageants et des améliorations supplémentaires peuvent être proposées et rapidement testées.

8.1 Directions futures

Une première approche serait sans doute d'optimiser le potentiel proposé dans [Kleinman et al., Submitted], car un potentiel composé uniquement d'un terme de contact et d'un terme d'accessibilité au solvant peut paraître trop simple au premier abord. Cependant, ce terme est suffisant pour tester de nouvelles améliorations du point de vue de la méthode d'optimisation, que l'on pourrait ensuite étendre à d'autres paramètres, comme par exemple ceux définis par [Kleinman et al., Submitted].

8.1.1 Affiner le terme d'interaction

Avant de s'intéresser à l'intégration de nouveaux termes dans le potentiel statistique, on peut d'abord essayer de raffiner ceux existants. La définition des classes d'accessibilité

au solvant ayant été intensivement étudiée (cf. chapitre 4), nous nous concentrerons plutôt ici sur le terme d'interaction.

8.1.1.1 Redéfinition des contacts

Comme nous l'avons vu dans le chapitre 2 (fig. 2.5), la définition des centres de contact et de la distance seuil pour laquelle on considère qu'un contact a lieu ont une grande influence sur la matrice de contact. De plus, dans notre formalisation actuelle, la matrice de contact est définie une fois pour toutes de sorte que les contacts d'un site i soient identiques, quel que soit l'acide aminé placé en ce site i. Il s'agit toutefois d'une simplification qui mérite d'être examinée de plus près. Nous avons en effet choisi de prendre le centre de masse des acides aminés comme centre de contact. En théorie, différents acides aminés que l'on teste en une position donnée n'auront pas leur centre de masse au même endroit (les acides aminés plus gros ayant leur centre de masse plus loin de la chaîne principale), ce qui montre que la nature des acides aminés en chaque site influence la matrice de contact (le point est illustré sommairement dans la figure 8.1).

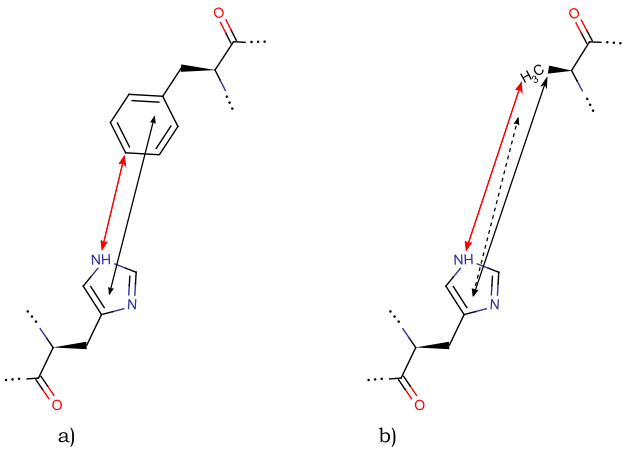


FIGURE 8.1 – Influence de la nature de l'acide aminé dans la définition des contacts (les pointillés indiquent la continuité de la chaîne principale). La distance entre les deux acides aminés entre les deux sites i (en haut) et j (en bas) est influencée par la nature des deux acides aminés en contact. a) site i : phénylalanine, site j : histidine b) site i : alanine, site j : histidine. Les flèches représentent des distances entre les acides aminés (en noir : entre les centres de masse, en rouge : entre les deux plus proches atomes).

La figure ci-dessus suggère également que cette influence sera encore plus importante quand l'on utilisera la définition "tous atomes" des contacts (cf. section 2.2.4.1). Et donc, même dans le contexte *coarse-grained* dans lequel nous avons travaillé, la description

structurale devrait idéalement être ajustée quand on remplace un acide aminé par un autre dans la séquence.

Une manière de procéder consisterait à définir, pour chaque paire de positions, non pas une seule variable prenant comme valeur 1 ou 0, mais un vecteur de 1 ou 0 associé à chaque paire possible d'acides aminés. Il s'agirait donc, lors de la procédure de pré-optimisation, de définir une nouvelle matrice de contact, de taille $210 \times N \times N$, au lieu de la matrice de taille $N \times N$, où N est la taille de la séquence [32].

Il serait sans doute complexe d'essayer de créer une telle matrice de contact vectorielle. Il faudrait également pouvoir placer, dans la procédure de pré-traitement, chaque paire d'acides aminés en chaque paire de sites, puis vérifier s'ils sont ou non en contact [Bastolla et al., 2001,Launay et al., 2007]. De plus, dans une telle procédure, le positionnement des chaînes latérales posera problème, et notamment au niveau des exclusions stériques. Une solution pourrait être d'utiliser le rotamère le plus fréquent [Launay et al., 2007], ou pourquoi pas, d'utiliser un programme de dynamique moléculaire pour optimiser cette chaîne latérale. Comme il s'agit ici de construire des matrices de contact, il est possible qu'une telle méthode donne de moins bonnes améliorations dans le cas de la définition des contacts que nous avons choisi (entre les centre de masses des chaînes latérales) que dans le cas d'une définition "tous-atomes" des contacts. Une telle définition pourrait par contre se révéler critique dans le cadre des potentiels de distance discrétisés.

8.1.1.2 Potentiel de distance

Le potentiel de distance discrétisée est le potentiel de distance le plus utilisé (cf. [Jones et al., 1992a, Sippl, 1993a] par exemple) : il s'agit de partitionner l'espace des distances entre deux sites en plusieurs classes, puis d'assigner une énergie de contact à chaque classe. Le potentiel de distance qui semble donner les meilleurs résultats dans le cadre que nous avons développé dans le chapitre 4 est présenté dans [Kleinman et al., Submitted]. Dans le chapitre 7, j'ai utilisé un potentiel de distance discrétisé utilisant trois matrices ($v_{seuil} = \{4,5\ ;\ 6,5\ ;\ 8,5\}$ Å).

Une alternative séduisante, mais plus complexe, est proposée par la formule de Lennard-Jones :

$$E_{distance}(lm) = \left(\frac{A_{lm}}{r_{lm}}\right)^{12} - \left(\frac{B_{lm}}{r_{lm}}\right)^{6}, \tag{8.1}$$

où l et m sont deux corps que l'on suppose en contact, et r_{lm} est la distance qui les sépare. Le premier terme représente la répulsion à faible distance, et le deuxième terme l'attraction

32. En pratique, la taille de la matrice de contact est creuse, et donc, peut être représentée de manière beaucoup plus compacte.

à moyenne distance. Lorsqu'on observe un clash stérique, l'énergie devient extrêmement élevée, et les corps se repoussent violemment. Dans la formule de Lennard-Jones, l et m sont originellement des atomes, mais on peut aussi appliquer un potentiel de Lennard-Jones aux pseudo-atomes de la représentation *coarse-grained* considérée. Si la formulation la plus commune du potentiel de Lennard-Jones est celle présentée à l'équation (8.1), également appelée potentiel $(6, 12)$ de Lennard-Jones, il existe des formulations $(4, 8)$, qui permettent notamment d'adoucir un peu le potentiel de répulsion.

Un tel potentiel est attrayant, car il permet notamment de représenter les clashes stériques avec un plus grand réalisme qu'un simple potentiel de distance. Cependant, il présente de nombreux désavantages, notamment à cause, justement, des instabilités numériques engendrées par le terme de répulsion en présence de clashes stériques. En effet, lorsque l'on souhaite optimiser un potentiel de distance en utilisant la formule de Lennard-Jones, on observe très rapidement des problèmes liés à ce terme de répulsion, qui cause un gradient engendrant des variations très fortes dans des valeurs des paramètres du potentiel, A_{lm} et B_{lm}. Comme cela se produit pour toutes les paires possibles de résidus (soit 420 paramètres), le paysage énergétique devient très accidenté. Et donc, pour obtenir les valeurs de potentiel optimal, la descente de gradient devra être faite avec un pas très petit, ce qui la rendra très lente. A noter que des instabilités de ce type risquent de poser également problème dans les applications phylogénétiques, en rendant le modèle très sensible aux clashes stériques.

L'utilisation de potentiels de Lennard-Jones soulève d'autre questions : en particulier, faut-il définir un seuil au delà duquel on ne considère pas que deux sites sont en contact ? Théoriquement, ce potentiel demanderait au contraire qu'aucune limite ne soit fixée pour la distance de contact entre deux sites, et en outre, la distance changera de manière importante selon l'acide aminé testé en chaque site. Enfin, quel sera le coût d'un tel potentiel dans le contexte des modèles d'évolution structurellement contraints ? En effet, le cadre dans lequel ce modèle est implémenté est déjà lourd et complexe, et demande à ce que le potentiel soit calculé rapidement.

Dans l'immédiat, on préfèrera donc conserver le potentiel de distance discrétisé, qui semble bien plus adapté à notre contexte, ou en tout cas d'utilisation plus simple.

8.1.2 Amélioration de l'approche par structures leurres

Bien que l'amélioration des différents termes du potentiel soit une piste de recherche intéressante, j'ai préféré, au sein de cette thèse, me focaliser sur le problème des méthodes d'optimisation, afin notamment de les comparer entre elles. La dernière approche présen-

tée dans ce manuscrit (chapitre 7), portant sur la prise en compte explicite de structures leurres pour évaluer la fonction score, m'intéresse plus particulièrement. Même si les premiers résultats sont modestes, cette approche mérite de plus amples réflexions, et des tests supplémentaires seraient extrêmement intéressants à mettre en œuvre. Dans ce qui suit, je détaille les points principaux que j'aurais souhaité développer si j'en avais eu le temps, et qui me semblent être les directions les plus prometteuses pour améliorer l'approche par structures leurres.

8.1.2.1 Pondération des structures leurres

Comme on l'a vu dans le chapitre 7, le terme de normalisation, $\ln Z_s$, peut-être approximé (à une constante additive près) par :

$$\begin{aligned} \ln Z_s &\simeq \langle E(s,c') \rangle_{c'} \\ &\simeq \sum_{c'} p(c'|s) E(s,c'). \end{aligned} \tag{8.2}$$

Cela signifie que les structures qui doivent avoir le plus de poids dans la détermination du $\ln Z_s$ sont les structures c', $c' \sim p(c'|s)$, qui sont considérées comme étant les plus stables par le potentiel. Il serait donc intéressant de constituer un jeu de données de structures leurres et de sélectionner parmi elles les plus importantes pour l'optimisation du potentiel, à chaque étape de la descente du gradient. On peut noter la similitude entre cette approche, et les méthodes développées par Thomas et Dill dans le cadre du *protein folding*, ainsi que notre approche par échantillonnage de Gibbs dans le cas du *protein design*. Cependant, un tel échantillonnage serait sans doute compliqué à mettre en œuvre à chaque étape de la descente de gradient. De plus, comme un tel échantillonnage devrait certainement se faire à l'aide de MCMC, on rajouterait une complexité supplémentaire, bien trop semblable à celle justement éliminée par l'utilisation de l'approximation issue du *leave-one-out*.

Une alternative à l'échantillonnage dans un ensemble de structures, serait de se définir un ensemble raisonnablement large, mais fini, de structures leurres, et de pondérer ces structures en fonction de leur ressemblance avec la structure native. Dans cette direction, la formulation de l'*overlap* moyen, utilisée notamment par Bastolla et collaborateurs, est une mesure naturelle de la distance entre deux repliements. Bastolla et collaborateurs avaient introduit l'*overlap* moyen dans une approche visant à optimiser un potentiel statistique. Spécifiquement, (voir par exemple [Bastolla et al., 1999]), l'on cherche à trouver

les potentiels qui maximisent l'*overlap* natif moyen :

$$\langle q(\tilde{c}, c') \rangle = \frac{q(\tilde{c}, c') \exp(-E(c', s))}{\sum_{c''} \exp(-E(c'', s))}, \tag{8.3}$$

où $\tilde{c}$ représente la matrice de contact native et

$$q(\tilde{c}, c') = \frac{N_{\tilde{c}c'}}{\max(N_{c'}, N_{\tilde{c}})}. \tag{8.4}$$

N'_c représente le nombre de contact dans la matrice de contact c' et $N_{\tilde{c}c'}$ le nombre de contact communs entre les deux matrices de contact. L'*overlap* tend vers 1 quand les deux matrices de contact sont très proches, et tend vers 0 quand elles sont très éloignées. Ainsi, utiliser le terme d'*overlap* moyen permet de pénaliser les matrices de contact des structures leurres les moins proches par rapport aux structures proches.

De la même manière, on peut donc utiliser la mesure offerte par l'overlap moyen pour introduire dans notre méthode un terme permettant de pondérer les structures leurres en fonction de leur distance à la structure de référence. Le fit statistique de schémas de pondération alternatifs pourront, encore une fois, être tous comparés au sein de notre cadre statistique de validation croisée. On peut même envisager de reproduire directement le critère défini par Bastolla et al dans leurs modèles structurellement contraints, à partir de l'overlap moyen. Ainsi, cela nous permettrait notamment de comparer l'efficacité de notre méthode et de celle de Bastolla et al, dans le même contexte du *protein design*.

8.1.2.2 Apprentissage séparé selon les classes de protéines

On remarquera que, dans l'optimisation de potentiels statistiques, on observe souvent une dépendance des potentiels aux tailles des protéines considérées. Au delà, les matrices de contact des protéines sont différentes selon leur taille [33]. Par exemple, une matrice de contact pour une protéine de 50 acides aminés est beaucoup plus compacte qu'une matrice de contact pour une protéine contenant 200 acides aminés. Et, lorsque l'on cherchera à prendre la structure de la grande protéine comme structure leurre pour la petite protéine, on observera bien moins de contacts que si l'on avait utilisé une autre petite protéine comme structure leurre.

Ainsi, une méthode d'apprentissage pourrait séparer les protéines en différentes classes selon leur taille. Plusieurs tests seront nécessaires pour déterminer si la taille des protéines est significativement importante pour déterminer le potentiel statistique. On peut envisager deux approches différentes : soit construire des ensembles de structures leurres

33. Les protéines contenant moins de 50 acides aminés sont notamment très particulières, et c'est pour cela que nous ne les avons pas considérées dans l'apprentissage de potentiel.

différentes pour chaque classe, soit définir un potentiel différent pour chaque classe. En principe, les mêmes forces sont censées s'appliquer pour toutes les protéines, et donc, définir des potentiels différents pour différentes classes de protéine peut sembler inadéquat. Mais puisque les potentiels statistiques sont des champs de force phénoménologiques, on peut supposer que leurs valeurs optimales pourront être différentes, en fonction de la taille des protéines. Si cela est le cas, il faudra aussi déterminer le nombre de classes optimal pour l'apprentissage. Au delà du *clustering* par taille des protéines, on peut également imaginer un clustering par type de protéines, et vérifier si la fonction de la protéine est une contrainte suffisante pour définir un potentiel statistique pour chaque classe.

8.1.2.3 D'autres méthodes de création de structures leurres

Les structures leurres obtenues par *threading* sont considérées comme de mauvaises structures pour la détermination de potentiels dans le cadre du protein folding. Elles sont en effet des compétiteurs trop peu performants, car trop différents de la structure native. Aussi, il existe plusieurs méthodes permettant d'obtenir des structures de meilleure qualité afin d'essayer de mieux entrer en compétition avec la structure native. Malheureusement, dans beaucoup de cas, il s'agit de faire appel à des programmes de dynamique moléculaire qui donnent lieu à des structures leurres qui ne sont pas forcément adaptées à notre contexte. Par exemple, les structures définies dans [Rajgaria et al., 2008] ne conviennent pas, comme nous l'avons vu dans le chapitre 7, car elles sont, à l'inverse des leurres obtenus par *threading*, trop proches de la structure native, pour en être suffisamment distinguées dans le contexte d'une description *coarse-grained*.

Cependant, il existe une méthode, définie notamment dans [Vendruscolo and Domany, 1998a], permettant de définir des structures dans l'espace des matrices de contact, qui conservent (théoriquement) une cohérence physique, tout en étant adapté au niveau de *coarse-graining* qui nous intéresse. Cette construction se fait en plusieurs étapes : à partir d'une matrice de contact, la procédure se déroule comme suit :

- Dynamique générale : au sein de la matrice de contact, plusieurs (gros) groupes de contact, sont déplacés dans la matrice (la matrice initiale et la matrice finale deviennent donc significativement différentes).
- Dynamique locale : des petits déplacements locaux des contacts sont effectués au sein de la matrice.
- Reconstruction : il s'agit de faire en sorte que cette nouvelle matrice retrouve une cohérence physique, à l'aide d'un algorithme défini dans [Vendruscolo et al., 1997].
- Raffinement : les structures sont raffinées à l'aide d'une fonction d'énergie.

En oubliant la dernière partie de l'optimisation, il serait ainsi possible de générer de nouvelles matrices de contact leurres, qu'il serait particulièrement intéressant de tester par rapport aux structures leurres obtenues par threading, dans un cadre d'optimisation de potentiels et de comparaison de modèles, tel que défini dans le chapitre 7.

8.1.2.4 D'autres structures leurres variationnelles

Nous avons vu, dans le chapitre 7, que l'on pouvait également construire des leurres dits virtuels, par le biais d'une méthode variationnelle. On peut imaginer plusieurs variations dans la définition des structures leurres pour les matrices de contact. Notons par la suite $\chi(ij)$ la structure leurre variationnelle entre i et j, telle que $0 \leq \chi(ij) \leq 1$. Dans la méthode présentée au chapitre 7, on fixait :

$$\chi(ij) = \begin{cases} 0 & \text{si } |j-i| > 50 \text{ ou si } |j-i| < 3 \\ \chi_{|j-i|} & \text{sinon} \end{cases}, \tag{8.5}$$

et on optimisait χ_d pour $d = \{0..50\}$ en même temps que les autres paramètres du potentiel. Cependant, le partitionnement de l'espace ici défini n'est peut-être pas le plus adapté. Par exemple, Gromiha et al définissent trois types d'interaction dépendant de la distance en acides aminés dans la séquence : *short-range* (séparés par un seul site au maximum), *medium-range* (séparés par trois ou quatre sites au maximum) et *long-range* [Gromiha and Selvaraj, 2004]. Le dernier type de contact fut par la suite partitionné en plusieurs classes(4-10, 11-20, 21-30, 31-40, 41-50), pour déterminer, selon les types de protéines, les pourcentages de contact entre deux sites. D'après eux, la distance générale à partir de laquelle les contacts entre deux sites devient négligeable est comprise entre 21 et 30 sites. D'un autre côté, Rossi et al posèrent que les matrices de contact leurres entre les deux sites i et j étaient égales à la même valeur si $|j-i| \geq 16$ [Rossi et al., 2001]. Ces partitionnements mériteraient d'être testés de manière plus systématique dans ce contexte.

8.1.3 Optimisation de potentiels statistiques dans un modèle d'évolution

Dans toutes les améliorations citées dans ce chapitre et les précédents, l'optimisation des potentiels est toujours décrite en dehors du modèle d'évolution lui-même. Dans un premier temps, nous avons même introduit des incohérences entre le modèle d'évolution et l'étape d'optimisation du potentiel, que nous avons du corriger en introduisant un terme correctif au potentiel, dans le contexte phylogénétique, pour pallier à la sur-représentation

de la partie mutationnelle (cf. chapitre 6). L'idéal, pour optimiser des potentiels statistiques pour les modèles d'évolution, serait d'optimiser les potentiels à l'intérieur même du modèle phylogénétique. Théoriquement, cela est possible, car la méthode d'optimisation et le modèle d'évolution sont décrits de manière à rester cohérents entre eux. Mais cela pose plusieurs problèmes pratiques.

Tout d'abord, le potentiel doit être optimisé sur un nombre de protéines suffisant, pour éviter les problèmes liés à un manque d'information, ou à des potentiels qui seraient trop spécifiques aux protéines du jeu d'apprentissage : il faut au moins 300 protéines partageant peu d'homologie pour que le potentiel puisse être généralisable. Étant donné la lourdeur computationnelle du MCMC phylogénétique, une telle approche demanderait d'effectuer un lourd travail de parallélisation des calculs, sur une grappe de calcul. Il faut donc se poser la question de savoir si un tel effort a des chances d'apporter de réelles améliorations.

De plus, le modèle d'évolution auquel nous nous intéressons est un modèle mutation/sélection, où les mutations sont appliquées sur les séquences nucléotidiques. Ainsi, il faut donc retrouver les séquences nucléotidiques correspondant aux protéines qui serviront à l'optimisation du potentiel. De plus, pour optimiser les potentiels dans le contexte du modèle d'évolution, il faut des données supplémentaires : pour chaque protéine, il faudra récupérer différentes séquences protéiques, issues de différents organismes, puis extraire leurs séquences nucléotidiques et enfin aligner ces différentes séquences nucléotidiques. On risque au passage, au cours de cette étape d'alignement à grande échelle, pour de nombreuses familles de structures, de mettre en évidence la faiblesse de l'hypothèse de constance de la structure le long de la phylogénie.

Une première étape dans cette direction pourrait être d'optimiser d'abord les potentiels à l'extérieur du modèle d'évolution, puis de raffiner les paramètres obtenus sur un jeu de données phylogénétique. En faisant tourner le modèle jusqu'à obtenir les distributions stationnaires des paramètres, il sera alors possible d'étudier ce modèle, par exemple en vérifiant les probabilités postérieures (*post predictive test*, [Rodrigue, 2007]), ou en comparant ce modèle à celui que nous avons présenté au chapitre 6. Ceci nous permettra d'évaluer s'il vaut mieux sacrifier à la précision et conserver l'approximation que nous avons faite (consistant à séparer la partie optimisation et l'application phylogénétique) ou s'il est indispensable de combiner les deux approches.

8.2 Applications

8.2.0.1 *Protein design* et *protein folding*

Le premier potentiel statistique proposé par Miyazawa et Jernigan, qui reste encore aujourd'hui le potentiel de référence, avait été créé dans une optique de *protein folding*. Par la suite, de nombreuses approches ont été proposées, utilisant la même approximation quasi-chimique, ou réfutant une telle approximation. Un modèle exact de treillis a même été créé [Thomas and Dill, 1996b] afin de pouvoir tester différentes méthodes d'obtention de potentiels. Les différentes méthodes testées sur ce treillis [Chiu and Goldstein, 2000] (ou sur des protéines réelles [Qiu and Elber, 2005]) montrent de très bons résultats lorsqu'il s'agit de retrouver un repliement donné pour une séquence fixée.

Au contraire, lorsqu'il s'agit de retrouver une (ou des) séquences dans un contexte de *protein design*, les méthodes et les potentiels existants ne sont pas extrêmement performants. Cela peut être notamment causé par le fait que le *protein design* est un problème plus complexe que le problème du *protein folding*. En effet, une approche de *protein design* inclut implicitement une approche de *protein folding* car il faut retrouver une séquence qui puisse se replier dans la bonne structure, et là où le *protein folding* implique une recherche dans l'espace des structures, le *protein design* demande une recherche dans l'espace conjoint des séquences et des structures. Une autre explication possible des faibles performances des méthodes proposées jusqu'à maintenant est que le problème du *protein design* est plus 'stringent' que le problème du *protein folding*, du moins dans sa version simplifiée sous forme de *threading*. En effet, il n'est pas si difficile de retrouver le repliement natif d'une séquence, parmi un ensemble finalement assez restreint de repliements disparates. Par contre, retrouver la séquence native parmi toutes les séquences possibles d'une taille donnée N est beaucoup plus difficile, étant donné la taille de l'espace de recherche (20^N).

8.2.0.2 Séquences ancestrales

Retrouver la séquence d'un gène d'intérêt telle qu'elle existait chez le dernier ancêtre commun d'un groupe taxonomique donné (par exemple, le dernier ancêtre commun des mammifères) est un rêve qui pourra sans doute devenir réalité dans les années qui viennent. Déjà, plusieurs méthodes permettent d'inférer des fragments de séquences ancestrales avec beaucoup d'exactitude [Bourque et al., 2004]. Ces séquences ancestrales peuvent alors être comparées aux séquences actuelles, par leurs compositions, leurs propriétés... Il est même possible d'essayer de retrouver la fonction originelle de la protéine. Par la reconstruction

efficaces de séquences ancestrales, il est alors possible de mieux comprendre l'évolution, son processus, et même d'inférer des propriétés fondamentales du développement et de la morphologie des organismes ancestraux.

Le modèle d'évolution soumis à des contraintes structurales, qui constitue le cadre dans lequel s'articule cette thèse, est particulièrement intéressant dans ce contexte. A l'aide des potentiels statistiques optimisés et des alignements de séquences, il nous est possible de déterminer, dans la phylogénie reconstruite, la séquence à chaque nœud de l'arbre. En réalité, ces séquences sont constamment échantillonnées au cours du MCMC phylogénétique, et sont donc un sous-produit immédiat de la méthode développée par Nicolas Rodrigue. Plus particulièrement, nous avions représenté dans la figure 5.5 le profil de la séquence ancestrale inférée par le modèle (que nous reproduisons une nouvelle fois dans la figure 8.2). Il serait également intéressant de comparer les différentes séquences

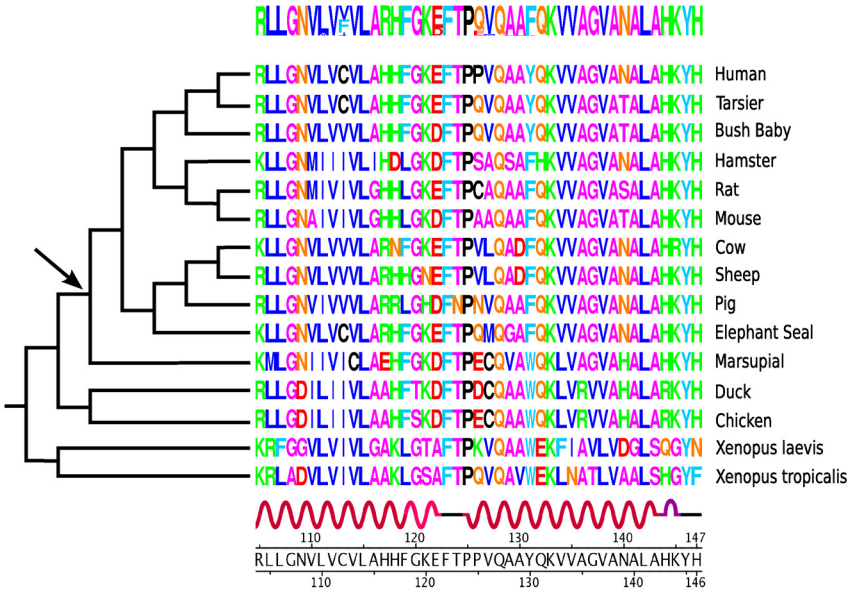


FIGURE 8.2 – Profil de la séquence inférée pour la séquence ancestrale de l'hémoglobine des mammifères [Bonnard et al., 2009]. La flèche indique le nœud correspondant.

ancestrales inférées à l'aide des différents potentiels, et de contraster les séquences obtenues sous des modèles structurellement contraints et des modèles phylogénétiques plus classiques.

8.2.0.3 Mesurer l'intensité de la sélection

On sait actuellement que l'hypothèse d'une vitesse d'évolution constante le long des sites et le long des branches d'une phylogénie n'est pas réaliste. Les modèles d'évolution phylogénétique les plus efficaces sont des modèles empiriques représentant l'hétérogénéité des vitesses d'évolution le long des sites par une loi gamma [Yang, 1993], ou en utilisant des modèles de mélange (voir par exemple [Le et al., 2008b]). Les variations entre sites sont d'ailleurs l'un des phénomènes que l'on a tenté d'expliquer (partiellement) à travers un modèle mécanistique dépendant de la structure.

D'un autre côté, l'on sait également que des variations des vitesses d'évolution sont causées par des fluctuations dans l'intensité de la sélection, qui est notamment causée par des variations de la taille des populations. Or, la taille de la population est un paramètre qui intervient dans les modèles structurellement contraints, soit directement (en utilisant le formalisme mutation/sélection présenté au chapitre 1), soit indirectement, via le paramètre β représentant la stringence de la sélection.

Plus spécifiquement, d'après le modèle que nous utilisons, la matrice de substitution entre deux codons c et c' est définie par :

$$R_{cc'} = \begin{cases} Q^{mut}_{bb'} \cdot e^{\beta(H(s|\gamma)-H(s'|\gamma))} & \text{si } a' \neq a \\ Q^{mut}_{bb'} & \text{si } a' = a \\ 0 & \text{si } c \text{ et } c' \text{ ne sont pas plus proches voisins} \end{cases}, \quad (8.6)$$

où $H(s|\gamma)$ est la fonction de sélection qui mesure l'adéquation entre la structure protéique et la fonction. Le paramètre β mesure l'intensité de la sélection. Si l'on décide de laisser ce paramètre s'adapter sur les données, et en particulier, en l'autorisant à varier le long de la phylogénie, alors, il nous permet de mesurer les variations de la stringence de la sélection, que l'on peut ensuite interpréter en termes de variations de la taille des populations au cours du temps : Si β augmente, alors la sélection augmente en intensité, signe que la taille de la population augmente. Au contraire, si β diminue, alors cela signifie que la taille de la population diminue. Et donc, mesurer β et sa variation le long des lignées serait extrêmement intéressant pour comprendre le processus évolutif au niveau de l'écologie des populations au cours de l'évolution.

Il serait également possible de reformuler le modèle d'une manière plus exacte, par rapport aux mécanismes de génétique des populations. En effet, on a vu dans le chapitre 1 que la probabilité de fixation d'un allèle pouvait être calculée de manière standard, en fonction du coefficient de sélection σ et de la taille de la population N ($p = \frac{2\sigma}{1-e^{-4N\sigma}}$), et que le taux de substitution était théoriquement proportionnel au taux de mutation multiplié par

la probabilité de fixation du mutant. En faisant l'hypothèse que le coefficient de sélection s'identifie à la variation de la fonction score, donc que $s = H(s|gamma) - H(s'|gamma)$, et en appliquant l'argument développé au chapitre 1, on peut alors construire le modèle d'évolution suivant :

$$R_{cc'} = \begin{cases} Q^{mut}_{bb'} \cdot \frac{4N(H(s|\gamma)-H(s'|\gamma))}{1-e^{-4N(H(s|gamma)-H(s'|gamma))}} & \text{si } a' \neq a \\ Q^{mut}_{bb'} & \text{si } a' = a \\ 0 & \text{si } c \text{ et } c' \text{ ne sont pas plus proches voisins} \end{cases}, \tag{8.7}$$

De cette manière on a construit un modèle explicitement dépendant de la taille de la population. On peut ensuite faire varier ce paramètre le long de la phylogénie.

8.3 Conclusion

D'une manière générale, cette thèse développe un cadre probabiliste systématique permettant de comparer différentes méthodes de création de potentiels statistiques, et d'obtenir des réponses claires sur les paramètres qui sont importants à prendre en compte dans une telle procédure d'optimisation, à l'aide de tests statistiques standards. J'ai présenté dans les deux premières sections de ce chapitre quelques améliorations rapides que l'on peut implémenter, et suggéré quelques pistes pour l'application d'un tel modèle (comprenant le modèle SC et la méthode d'optimisation).

Cependant, on peut se poser la question, finalement, de pourquoi optimiser des potentiels statistiques dans le contexte du *protein design*. On a vu dans la section précédente que c'est un domaine plus stringent que le *protein folding*, mais est-ce réellement important ? Théoriquement, une même fonction devrait pouvoir être utilisée dans le cadre du *protein design* et du *protein folding*. Cependant, à cause des diverses approximations induites par la forme du potentiel statistique et les domaines de recherche (ensemble des séquences, ensemble des structures), il est probable qu'un potentiel optimisé dans une optique ne soit pas optimal pour l'autre. On a vu par exemple dans les chapitres 4 et 6 que le potentiel que nous avons optimisé est bien meilleur que le traditionnel potentiel de Miyazawa et Jernigan, lorsqu'on utilise les mêmes paramètres (inclusion de terme de normalisation et du terme de correction mutationnel). Cependant, une telle comparaison n'est peut-être pas la plus informative, dans la mesure où la différence de fit pourrait être due à bien d'autres causes (utilisation par Miyazawa et Jernigan d'une base de données bien plus petite que nous, ou problèmes liés à l'utilisation de l'approximation quasi-chimique). Il faudrait donc faire des tests plus poussés, en ré-optimisant des potentiels pour le *protein*

folding, possiblement en utilisant des méthodes analogues à celle proposée par Thomas et Dill, puis en testant ces différents potentiels à l'aide des différentes méthodes que j'ai présentées ici.

Globalement, les résultats présentés montrent que l'amélioration des potentiels semble devoir être conditionnée à une description structurale de plus en plus poussée. Cependant, s'il est vrai que les potentiels présentés notamment dans [Kleinman et al., Submitted] proposent des améliorations tangibles, cela se fait au détriment des avantages computationnels proposés par l'approche des potentiels statistiques. Il en est de même pour l'approche utilisant les structures leurres dans la procédure d'optimisation et dans le modèle d'évolution. De plus, si la description de plus en plus sophistiquée permet d'introduire des termes qui miment les champs de force semi-empiriques, tout en n'atteignant pas leur complexité, ils n'atteignent pas la précision offerte par ces champs de force.

Au delà, le modèle d'évolution soumis à des contraintes structurales prend comme hypothèse que l'évolution est principalement subordonnée aux structures. Cependant, les simples modèles à omega fournissent toujours de meilleurs résultats que la formulation la plus élaborée du potentiel [Kleinman et al., Submitted]. Puisque ce paramètre est utilisé notamment pour dire si la sélection est positive ou négative, cela peut soit signifier que la sélection par la structure tridimensionnelle n'est finalement pas si importante que cela, ou bien que notre manière de définir le modèle n'est pas la plus adaptée. Au delà même de l'utilisation de potentiels statistiques, maintenir la structure exactement constante le long de l'évolution n'est peut-être pas non plus une bonne approximation. Après tout, même si les structures évoluent peu comparées aux séquences [Chothia and Lesk, 1986], elles varient au cours du temps. Jusqu'à présent, nous nous sommes contentés d'espérer que les potentiel de contact et de distances discrétisées devraient normalement nous affranchir d'un tel souci, la description structurale étant suffisamment grossière pour rendre la méthode insensible à d'éventuelles variations de structures au cours du temps. Toutefois, c'est un problème réel, qu'il s'agirait de creuser.

Malgré tout, la définition du cadre statistique permettant le test systématique de différentes méthodes d'optimisation dans un même contexte et avec les mêmes paramètres est intéressant. Les résultats peuvent être surprenants (comme par exemple les structures leurres qui permettent d'obtenir une amélioration réelle, mais modeste), mais ils permettent cependant d'affirmer ou de réfuter les hypothèses proposées par le modèle. Si les améliorations proposées dans cette thèse sont modestes, le cadre statistique offert est robuste et systématique, tout en permettant l'inclusion de nouvelles formes de potentiel ou de nouvelles techniques d'optimisation.

Conclusions

Les travaux réalisés dans cette thèse visent à optimiser des potentiels statistiques pour un modèle phylogénétique soumis à des contraintes structurelles. À l'articulation entre la phylogénie et l'étude structurale des protéines, le but est aussi de fournir un modèle entièrement probabiliste liant ces deux domaines. Au sein d'un modèle d'évolution SC bayésien, nous avons tout d'abord défini une première version de notre modèle probabiliste (chapitre 4), qui fut par la suite partiellement reformulé (chapitre 6) afin de le définir entièrement dans le cadre phylogénétique probabiliste. Plusieurs améliorations computationnelles (chapitre 5) permettent l'exploration de nombreuses formes de potentiels, et de critères d'optimisation de potentiels (chapitre 7), qui, s'ils donnent des résultats intéressants, mériteraient d'être plus amplement explorées (chapitre 8).

Le cadre statistique du modèle phylogénétique est très complexe et très demandant d'un point de vue computationnel, ce qui ne nous permet pour le moment que de travailler à topologie fixée. Cependant, de nombreuses méthodes de reconstruction phylogénétiques permettent d'inférer des topologies correctes, et il nous est donc possible de nous concentrer sur les processus mutationnel et sélectif sous-jacents, exprimés, entre autres, par les longueurs de branches. A travers ce modèle phylogénétique et les potentiels optimisés, il est possible de déterminer qu'il existe une contrainte évolutive sur les séquences nucléotidiques, liée à la structure de la protéine codée, et il est possible de déterminer, au sein de modèle choisi, si les potentiels optimisés sont effectivement valides, par l'observation du paramètre β, lors du calcul du facteur de Bayes (entre le modèle purement mutationel et le modèle SC). En effet, plus ce β sera proche de 0.5, plus on observera une consistance entre le modèle phylogénétique et le modèle d'optimisation.

Au delà de l'optimisation de ces potentiels statistiques et de la paramétrisation du modèle phylogénétique en fonction des contraintes structurelles, il est possible de définir bien d'autres fonctions mesurant l'adéquation d'une séquence protéique à son environnement, car le modèle est définie de manière entièrement probabiliste. La matrice de substitution du modèle phylogénétique est décrite par une matrice de mutation et une probabilité de fixation, qui ne dépend que des séquences protéiques et du contexte. La méthode d'opti-

misation ne est basée sur la méthode du maximum de vraisemblance, et donc, n'importe quelle propriété de la séquence protéique pouvant être décrite en termes probabiliste pourrait être une nouvelle cible à étudier. Par exemple, on pourrait introduire la relation entre la protéine et son ligand, et définir des potentiels statistiques (sans aucune valeur thermodynamique) pour les sites constituant le site actif. On pourrait également introduire une fonction qui dépendrait de l'environnement thermique des protéines considérées, afin de considérer l'influence de celle-ci sur l'évolution des protéines.

Mieux encore, ce modèle phylogénétique peut être modifié pour introduire des notions de génétique des populations [Choi et al., 2008], afin de relier de manière mécanistique les trois composantes de l'évolution (mutation, sélection et dérive génétique). Un tel modèle pourrait alors être utilisé pour générer des séquences ancestrales [Williams et al., 2006]. Cela pourrait avoir différents intérêts. L'on peut souhaiter connaître un profil de séquence pour un ancêtre lointain, et essayer de comprendre, à l'aide de la composition de la séquence, dans quel type de milieu il vivait.

D'autres applications moins théoriques peuvent être proposées pour ces séquences ancestrales. Utiliser un tel modèle structurellement contraint, utilisé dans un contexte de protein design pourrait permettre de réduire le nombre de séquences possibles pour une protéine ancestrale d'intérêt agronomique, par exemple. En outre, ces modèles peuvent également être utilisés pour essayer de prédire la fonction que pourrait avoir un protéine, ou de quelle fonction elle dérive, puisque les fonctions des protéines sont intrinsèquement liées à leurs structures tridimensionnelles. Et, pourquoi pas, être utilisés pour aider à la génération d'une protéine inexistante, mais qui répondrait à un besoin actuel, comme par exemple la génération d'un nouveau médicament ?

Annexe A

Liste des abbréviations

ADN : Acide DésoxyriboNucléique
ARN : Acide RiboNucléique
GLW : Méthode d'optimisation du Z-score par Goldstein, Luthey-Schulten, and Wolynes
GS : Echantillonnage de Gibbs
GY : Goldman et Yang
MH : Metropolis-Hasting
MCMC : Chaînes de Markov Monte Carlo
MG : Muse et Gaut
ML : Maximum de Vraisemblance
MS : Méthode d'optimisation du Z-score par Mirny et Shakhnovich
PDB : Protein Data Bank
REM : *Random Energy Model*
SC : soumis à des contraintes structurales

Annexe B

Développement du cadre statistique

B.1 Fichier additionnel 1

Extensive definition of accessibility classes

Solvent accessibility classes were defined in quantiles, i.e. each class has the same number of amino acids.

Upper limits (in % of solvent exposure relative to the Ala-X-Ala tripeptide) of the solvent accessibility classes used in the study, for each D :

2 classes : 21.3 ; 199.3 ;

4 classes : 3.5 ; 21.3 ; 48.1 ; 199.3 ;

6 classes : 0.9 ; 7.8 ; 21.3 ; 38.6 ; 58.5 ; 199.3 ;

8 classes : 0.3 ; 3.5 ; 10.6 ; 21.3 ; 34.1 ; 48.1 ; 64.8 ; 199.3 ;

10 classes : 0.2 ; 1.7 ; 5.9 ; 12.6 ; 21.3 ; 31.4 ; 42.3 ; 54.1 ; 69.3 ; 199.3 ;

12 classes : 0.1 ; 0.9 ; 3.5 ; 7.8 ; 13.9 ; 21.3 ; 29.7 ; 38.6 ; 48.1 ; 58.5 ; 72.7 ; 199.3 ;

14 classes : 0.1 ; 0.6 ; 2.2 ; 5.1 ; 9.4 ; 14.9 ; 21.3 ; 28.4 ; 36 ; 43.9 ; 52.3 ; 61.9 ; 75.4 ; 199.3 ;

16 classes : 0.1 ; 0.3 ; 1.4 ; 3.5 ; 6.6 ; 10.6 ; 15.6 ; 21.3 ; 27.5 ; 34.1 ; 41 ; 48.1 ; 55.7 ; 64.8 ; 77.5 ; 199.3 ;

18 classes : 0.1 ; 0.2 ; 0.9 ; 2.5 ; 4.8 ; 7.8 ; 11.7 ; 16.3 ; 21.3 ; 26.8 ; 32.6 ; 38.6 ; 44.8 ; 51.4 ; 58.5 ; 67.2 ; 79.2 ; 199.3 ;

20 classes : 0.1 ; 0.2 ; 0.7 ; 1.7 ; 3.5 ; 5.9 ; 8.9 ; 12.6 ; 16.7 ; 21.3 ; 26.2 ; 31.4 ; 36.8 ; 42.3 ; 48.1 ; 54.1 ; 60.8 ; 69.3 ; 80.8 ; 199.3 ;

B.2 Fichier additionnel 2

Data set DS1 - List of PDB identifiers of proteins used

Data set DS1

Proteins are named by their PDB id + chain identifier. When there is no chain identifier, the last character is '0'

1W2WB 1LYCA 1JO0A 1XDZA 1OQJA 1CSN0 1MG7A 1MNAA 1W2YA 1BEA0 1D2OA 1DP4A 1GVNA 1VLRA 1J98A 1R7LA 1VLS0 1MK0A 1IRQA 1GL2C 1WPCA 3CHBD 1QU9A 1M55A 1W6SA 1BB1B 1OQQA 1ALVA 1GL4A 1LK9A 1BEHA 1CT5A 1D2VC 1RIYA 1MK4A 1DI2A 1VICA 1VPMA 1GSA0 1VLYA 1J27A 1HDKA 1MUWA 1J1YA 1UQ5A 1NNHA 1AQ0A 1R0MA 1PG4A 1SBP0 1T0BA 1SFDA 1JH6A 1DWKA 1XB3A 1PFVA 1BX70 1VPSA 1ET1A 1OJQA 1PG6A 1V7BA 1NUUA 1Y71A 1C3CA 1F32A 1N1FA 1CM4B 1GOTG 1CLVI 1ELKA 1QFTA 1R0RI 1UMGA 1UJ2A 1VMBA 3EZMA 1JL0A 1RJDA 1T0HA 1U7LA 1FI2A 1GP6A 1TFE0 1N1JA 1XEOA 1JL1A 1UBKL 1R0UA 1CPO0 1M1QA 1RQPA 1R0VA 1AMF0 1VMEA 1ZPDA 1SBYA 1XPPA 1Y6ZA 1UIXA 1CTF0 1N97A 1KA1A 1TQGA 1S12A 1GWEA 1FLMA 1NRJA 1SN9A 1H6FA 1LVK0 1Q8IA 1OO0A 1OO0B 1T0PB 1EB6A 1KZFA 1NRLC 1PJXA 1MGTA 1JDPH 1W44A 1TN6A 1NKD0 1XEWX 1F74A 1JSDB 1M1ZA 1SQSA 1SG4A 1OZ2A 1U4BA 1CMCA 1H6LA 1T0TV 1Y08A 1IE9A 1CIPA 1SR4B 1NH2C 1GWMA 1VC1A 4BCL0 1WIWA 1G73A 1V84A 1EF1C 1P1MA 1SQWA 1MZBA 1C7KA 1OGOX 1F00I 1JZTA 1HT6A 1IHSI 1PO5A 1QREA 1XJ4A 1G6XA 1MOGA 1QCSA 1V7ZA 1QNRA 1L2HA 1MZGA 1D3VA 1EEXG 1N62B 1WC3A 1J34C 1RYLA 1EF8A 1DMHA 1TJOA 1GX5A 1N5UA 1KP6A 1W0NA 1K6FA 1RV9A 1NA3A 1IWLA 1IAPA 1PGS0 1NZEA 1RYOA 1RO0A 1XFFA 1IWMA 1W0PA 1PKHA 1N9LA 1GPPA 1VQQA 1PVGA 1SJYA 1D7PM 1DQAA 1J6OA 1A4MA 1B12A 1AJJ0 1Y80A 1E7LA 1EU1A 1N2EA 1T1DA 1XMTA 1NZJA 1A8D0 1QVEA 1K2XB 1DC1A 1KWGA 1H03P 1DQEA 1Y0KA 1S5AA 1VFYA 1D4AA 1HQ0A 1FJ2A 1CQMA 1R1TA 1PA1A 1FTRA 1DXRC 1L6KA 1VYBA 1KHYA 1FIUA 1GQ8A 1BQUA 1MHNA 1XQOA 1N6AA 1UV4A 1V54J 1R5LA 1S21A 1PL3A 1A8L0 1XFSA 1JB7B 1DY5A 2IGD0 1UYLA 1W4RA 1SZ7A 1VK4A 1V58A 1QL0A 1A8O0 1GU2A 1VJVA 1JIDA 1VYIA 1K77A 1UCSA 1DQPA 1EC7A 1SVFA 1ARB0 1IQ5B 1FC3A 1GJ7A 1NZYA 1QA7A 1JETA 1HTRP 1GXMA 1T1VA 1SRVA 1WJXA 1RW1A 1HBNC 1SVMA 1YNBA 1OI0A 1U24A 1DQZA 1B93A 1Q9UA 1UKFA 1LM8V 1XYIA 1ITUA 1NAR0 1PZXA 1NSZA 1W96A 1GMXA 1BYRA 1IXKA 1LF2A 1SZHA 1T9HA 1J3WA 1PI1A 1FJJA 1UKKA 1VKFA 1FG7A 1EYEA 1Q33A 1VKIA 1G4MA 1CCWA 1XGKA 1CNV0 1S2KA 1WZ8A 1OX0A 1PSWA 1Q2WA 1UZEA 1VKMA 1QWGA 1DYPA 1VKNA 1AOCA 1U9LA 1NWZA 1MF7A 1BRT0 1IUAA 1HQSA

1SACA 1RHS0 1UOLA 1J0PA 1IFRA 1FNLA 1DD9A 1V2BA 1LU4A 1LXJA 1WOCA 1USCA 1T6CA 1QWOA 1Q74A 1K0MA 1K7WA 1TOVA 1PX5A 1AOL0 1WKQA 1K87A 1SSQA 1TP6A 2BF9A 1QAUA 1M48A 1TZVA 1M0DA 1PWXA 1QB5D 1WDJA 1TAFA 1MML0 1DKIA 1F1MA 1CZAN 1MU5A 1MIXA 1GUXB 1LBU0 2PGD0 1TZYD 1D9CA 1GD0A 1BDMA 1U6DX 1X9IA 1ICFI 1PPRM 1LC5A 1VHH0 1R6WA 1XOFB 1W2FA 1BDO0 1RWRA 2BBKL 1H8PA 1JFXA 1HUW0 256BA 1NU4A 1UZXA 1YOCA 1JR7A 1D5TA 1GNUA 1EGWA 1MXRA 1Y9IA 1LFPA 1PTMA 1SAUA 1N7SB 1VHNA 1UT1A 1CHD0 1INLA 1T6SA 1REQB 3TDT0 1MFMA 1RIFA 1IJYA 1G5HA 1JV1A 1H97A 1T6UA 1JNIA 1UEBA 2PVBA 1Q7FA 1EKQA 1L8BA 1SX5A 1I1JA 1XZOA 1R7AA 1EKRA 1CZPA 1O82A 1O7SA 1GVEA 1F60A 1LR5A 1N7ZA 1JUVA 4UBPA 1K8KF 1EZGA 1HZ6A 1V2ZA 1T79B 1UASA 1MUGA 1ON2A 1PU6A 1MG4A 1NFP0 1P0HA 1EL6A
END

B.3 Fichier additionnel 3

Data set DS2 - List of PDB identifiers of proteins used

Data set DS2

Proteins are named by their PDB id + chain identifier. When there is no chain identifier, the last character is '0'

1K1EA 1I1QB 1JNRA 1COZA 1KNMA 1GDEA 1N0WB 1V73A 1KGDA 7HBIA 1KV7A 1U00A 1G5TA 1N0XP 1UAYA 1JYSA 1IGQA 2BJIA 1E2KA 1XZZA 1QTXB 1OXXK 1QXMA 1UWWA 1BB1A 1RU4A 1PBYB 1PBYC 1K8WA 1FDR0 1UTEA 1X6OA 1LV7A 1WEHA 1XSVA 1PN9A 2NACA 1XEDA 1U7BB 1V3HA 1I60A 1WM3A 1VIE0 1U09A 1NJRA 1OQVA 1HW1A 1W78A 1D2ZA 1V05A 1IKOP 1BX4A 1Q0RA 1OFZA 1IKPA 1X74A 1YLLA 2ARCA 1VPRA 1M93B 1U7GA 1T7MB 1P4OA 1Q0UA 1QXYA 1O54A 1QQP1 1QQP2 1WPNA 1QQP4 1VQ3A 1AIL0 1WPOA 1LO7A 1U7IA 1MCTI 1YT8A 2BK7B 1DSZA 1WER0 1Q16A 1TXJA 1FLEI 1VIMA 1U7KA 1HE1A 1R12A 1NUYA 1DPGA 1W07A 1SXRA 1T7RA 1HP1A 1MKAA 1XLYA 1IA9A 1G6GA 1SBXA 1S7ZA 1RY9A 2MLTA 1QNAA 1NZ0A 1M22A 1KS8A 1U7PA 1VMGA 1FT5A 1LKKA 1L5OA 2PSPA 1TC5A 1NRJB 2NLRA 1NYTA 1BTN0 2DNJA 1JKXA 1S14A 1NNWA 1H32A 1H2SB 1ELUA 1MVFD 1S4KA 1FM0D 1T4FM 1JSDA 1EAYC 1TUAA 1JHGA 1WMHB 1H6KA 1P1JA 1DEUA 1KOLA 1MDC0 1SR4C 1Y7BA 1ATZA 1JOSA 1UC7A 1L9LA 1I2TA 1R4XA 1V4AA 1U0SA 1SR8A 1YTLA 1MOF0 1VFJA 1H70A 1TCA0 1SR9A 1UXY0 1J2RA 1RUTX 1GPIA 1M5WA 1CFB0 1KSHB 1WMQA 1R1GA 1XIWA 1GX3A 1I39A 1N62C 1DMGA 1XQAA 1ZIN0 1AY7B 1JHSA 1S1DA 1UQTA 1YBKA

1NOGA 1SJWA 1UUJA 2SAK0 1Y0EA 1MSC0 1P9AG 1R1MA 2PTD0 1UFYA 1T56A
1KSOA 1UCDA 1G3KA 1WMXA 1OVNA 1DJ8A 1XFIA 1E0BA 1VQSA 1O9IA 1Y0HA
1Y7RA 1HX6A 1KWFA 1HM9A 1O66A 1D7UA 1QS1A 1POC0 1PRXA 1V8EA 1J73A
1P9GA 1F3VA 1PVMA 1TUWA 1SD4A 1GTKA 1DQGA 1Q9BA 1QH5A 1JM1A 1KHXA
1HQ1A 2BOPA 1EFDN 1NE2A 1AK00 1S5DA 1EU8A 1DBWA 1P5VB 1XG0A 1J77A
1SGMA 1V54D 1V54I 1R26A 1V54M 1JEKA 1VCHA 1DFMA 1CQQA 1RH40 1RKIA
1J79A 1IX9A 1T5BA 1YBZA 1L6PA 1KPGA 1SYYA 1YN9A 1UCRA 1NOX0 1QKRA
1N2SA 1R9FA 1GU4A 1EY4A 1RGZA 1AYL0 1ROCA 1GTVA 1E4FT 1IQ6A 1S29A
1M6PA 1GU7A 1KICA 1VGGA 1L6XB 1BGC0 1GBS0 1POT0 1Q2HA 1TVGA 1N2ZA
1GXQA 1H80A 1UGPA 1UGPB 1VD5A 1RZHH 1RZHM 1QW9A 1H0HB 1IU1A 1C1DA
1K07A 1ECA0 1MPGA 1D4TA 1S5UA 1OI2A 1XCLA 2HRVA 1U5KA 1QSGA 1K7CA
1PWBA 1VKCA 1ITVA 1WCV1 1D4XG 1X91A 1FR2B 1TS9A 1AGQA 1L3LA 1V5IB
1W1OA 1JU2A 1US5A 1OWLA 1PSRA 2MCM0 1JU3A 1RSGA 2BM3A 1R9WA 1JIWI
1JFBA 1C96A 1WKCA 1L3PA 1SHEA 1JIXA 1JBOA 1VH5A 1VKKA 1YD0A 1ROWA
1C8ZA 1NWWA 1DCS0 1NM8A 1X99A 1UVQC 1DD3C 1V9FA 1A620 1UW1A 1TVXA
1D8WA 1IJBA 1E1HA 1S2OA 1NIGA 1J4AA 1SDWA 1QSUA 1THFD 1VKPA 1GQZA
1M7GA 1VL1A 1NTHA 1JY2P 1JFLA 1JMVA 1P3CA 1I0VA 1H8EH 1H8EI 1RWHA
1WDDS 1QWNA 1C1YB 1UHEB 1M45B 1H4RA 6RLXA 6RLXB 1O06A 1I4MA 1OTFA
1VL5A 1G4YB 1JUBA 1USEA 1KMTA 1RLMA 1VKWA 1EZ3A 1BKRA 1B65A 1MJ4A
1NPYA 1S2XA 1B66A 1AVWB 1USGA 1JG1A 1LTZA 1FD3A 1DS1A 1WA5B 1UDVA
1GK9A 1XDFA 1XKPA 1GCQC 1DGWX 1GV9A 1T2WA 1C2AA 1MTYD 1XKRA
2BJ4A 1CVRA 1MC2A 1OXDA 1NTVA 1OIS0 1UI0A 1VP4A 1T6OB 1CDLE 1GO3F
1DVOA 1I58A 1C9OA 1JNDA 1YDGA 1VLAA 1UAIA 1ES5A 1N7SC 1XDNA 1KR4A
1M7YA 1HYP0 1CL8A 1B9WA 1EVLA 1Y66A 1NQEA 1TT8A 1WDVA 1V33A 3THIA
1JK3A 1VZYA 1O7QA 1SENA 1PBJA 1PMI0 1PMHX 1H99A 1A6M0 1IO0A 1LFWA
1FKMA 1DZKA 1VHTA 1WHO0 1GVFA 1VI4A 1FS7A 1KNGA 2TNFA 1K8KD 1K8KE
1VHVA 1UEHA 1A6Q0 1V6PA 1ON3A 1GVJA
END

B.4 Fichier additionnel 4

Multiple sequence alignment for sequence logos of figure 5. Multiple sequence alignment (Clustal format) used to generate sequence logos of figure 5.

```
CLUSTAL W (1.8) multiple sequence alignment

1gdeA           ALSDRLELVSASEIRKLFDIAAGMKDVISLGIGEPDFDTPQHIKEYAKEALDKGLTHYGP
Q8Y0E8          -LSSKVSTLAPSPIQRTREAREAGRDVVDLTLGEPDFATPEHICEAARRAIADGLTKYTP
Q9BVY5          ------------------------------------------------------------
Q8RAK7          -LSQNALQITPSMTLEITAKARQLKDVIDFGVGEPDFDTPDYIKEAAIEAIKKGYTKYTP
AAB2_RHIME      --ASRISSIGVSEILKIGARAAAMKPVIILGAGEPDFDTPDHVKQAASDAIHRGETKYTA
Q55453          ----RMDLVQSPVIPVVGQWIVDSPGTISLGQGVAFYPPPNEVAVAVRESLETPLHQYQP
AATB_RHIME      --ASRISSIGVSEILKIGARAAAMKPVIILGAGEPDFDTPDHVKQAASDAIHRGETKYTA
Q8YTF2          --ADRIQQLPPYVFARLDELKAKARDLIDLGMGNPDGATPQPVVDAAIQALQDPKNHYPP
Q9KC79          ELAKRVSTLTPSSTLAITAKAKELKDVIGLGAGEPDFNTPQYIIDAAVRSMEEGHTKYTP
Q92S71          SLSPRATAAPESGIVEVVNYARGREGLIPLWVGEGDLPTPDFVSRAAADALMAGETFYTW
Q93703          ---------TVNPVRKIADACAVPKKVIKLHLGDPSVGGKLPPSEIAVQAMHESVSGYGP
Q8YMS6          -LAARVSQVTPSITLAIAAKAKAMKDVCSFSAGEPDFDTPAHIKAAAAKALDEGKTKYGA
Q9HV76          SYSARSRAIEPFHVMALLARANELHDVIHLEIGEPDFTTAEPIVEAGRAALAAGHTRYTA
Q8U3E6          -MAHRIGVIQRSKIRELFERASKMENVISLGIGEPDFDTPNNIKEAAKRALDEGWTHYTP
Q9HUI9          ------------------------EEILLLSVGDPDFDTPAPIVQAAIDSLLAGNTHYAD
Q8ZVJ5          -FSPRIGALRESPTRKIDELREKLRDVILLSTGQPSIPPPREVREALGELLKVDTMGYTP
AAT_METEX       ------PYVFEQVNRIKAAARARGADIIDLGMGNPDLDAPRHVIEKLVETAGKPRTRYSA
Q9ZLG5          PYSSKVQSLSESATIAISTLAKELKDILSFSAGEPDFDTPQAIKDAAIKALNDGFTKYTP
AAT1_METJA      -ISSRCKNIKPSAIREIFNLATS--DCINLGIGEPDFDTPKHIIEAAKRALDEGKTHYSP
Q8XJ54          -MNKNVIGIEISGIRKFYNEVVKFPEAISLTLGQPDFPVPEKVKEAMIRAIEEGKTTYTA
Q8UFR3          -----MSKRSAVEPFHAMDILAEARPVISMAVGQPSHPAPKASLAAAQEALKHGRIGYTD
Q9W6U2          --ANRTNGIDKNVWVEFTQLAAA-YSKVNLGQGFPDFAPPKFVQEAFCHALNEGMHQYTR
Q9K7L1          -LSTRVQSIQPSGIRRFFDLASKMENIISLGVGEPDFVTPWNVREASISSMERGFTAYSA
O66737          ---DRLEKVSPFIVMDILAQAQKYEDVVHMEIGEPDLEPSPKVMEALERAVKEKTFFYTP
Q9Y9P0          -IQERVVEIEGETAFAYLAVARKLRRVISFGIGQPDFPTPHHIREAAKKALDEGFTGYTE
Q99V44          SLNSNSKYLRAPSIRQFSNRMNNLDDCVNLTIGQPDFPMPDVVKKAYIDAINNDKTSYSH
Q8RA61          -LSQNALQITPSMTLEITAKAKQLKDVIDFGVGEPDFDTPSYIKEAAIDAIKRGYTKYTP
PATA_BACSU      -LNPKAREIEISGIRKFSNLVAQHEDVISLTIGQPDFFTPHHVKAAAKKAIDENVTSYTP
Q98NB8          SLSRRGNVEPFHAMDVLAELKAQGVPVISMAVGQPSDPAPVRVRAAAARALQDGRIGYTD
Q97PQ9          -LSNRVESVTLAAGARAKALKAEGRDILSLTLGEPDFTTPKNIQDAAIASIRDGRSFYTV
Q9KAU1          AKEIEQNLTQSSWIRKMFEEGQRLKNVFDFSLGNPIIDPPAQLRAYANAPIQGG-HSYIP
Q9V0G5          ----SIEYAIRDVVLPARELEKKGIKVIRLNIGDPDFQPPEHMKEAYCKAIKEGHNYYGD
AAT_PYRKO       ------------NVYEFFNRINEVRPESRLDAGQPDIPVRREIIEEAVESLRRGETGYTS
Q9L0L5          -VSARVGAISESATLAVDAKAKALKPVIGFGAGEPDFPTPDYIVEAAVEAKNPKFHRYTP
Q979X6          -LSQRLQYVNESATVSASNYVEKLKKVYNFGIGEPDFTTPQHIIEYAFEMAKEGKTHYTP
Q8RCV4          -IAKRAKAIEISTIRYFFNMAREVEGAISLAIGEPDFITPEHIRNAAKKALDEGMTGYTV
AAT_SYNY3       -LTQRVSQVVPSITLEITAKAKAMRDVLSFTAGEPDFTTPPHIVEAAKLALDEGKTRYGP
Q8XJT3          -LSRKAQNIGASLTLALTAKAGELKDVVSFGVGEPDFNTPKNIIEAATRAMEEGKTKYTA
Q97KB8          -FSNRVHNMDFSPIRKLVPLSKAAEDVYHLNIGQPDVKTPDTFFQGI-TNYNEKIVKYSD
Q8ZDK4          -LDNVCYDIRGPVLKEAKRLEEEGNKVLKLNIGNPAFDAPDEILVDVIRNLPTAQ-GYCD
O59096          ALSDRLELVSASEIRKLFDIAAGMKDVISLGIGEPDFDTPQHIKEYAKEALDKGLTHYGP
Q98AR6          -LAHRTNLFGTSGTAAARAAAKAAKEIIDLTAGEIWSELAPTIRDGAIDAINKGVNRYTD
Q8TT07          -LSKQSEDIPPFYVMEVLESAKELRHIIHLEVGEPDFPTAPHICEAACAAIGKGLTKYTH
Q9R096          -LHQSLTMTKRLQARRLDGIDQNLYDVVNLGQGFPDFSPPDFATQAFQQATSGNLNQYTR
Q9RNK6          -LWQPMMINMLSVLSVMIELKSKGVDIITLGAGEPDFETPEFIKEAAIQAIHDGKTRYTN
```

```
Q9SIE1          SLSPRVQSLKPSKTMVITDLAATLVPVIRLAAGEPDFDTPKVVAEAGINAIREGFTRYTL
YD91_METJA      ------------------------ELIDMGVGEPDMADPEVIRVLCEEAKKWENRGYAD
Q9CPI6          --SDKLEHVCYDIRGPVHKEALRLHKILKLNIGNPAPFGFEAPDEILVDVIRNLPTGYCD
Q8U097          ------------NVYELFNKINEVKPEIRLDAGQPDIPVANEIIEEAINSLKRGETGYVS
Q9CEK7          ------ESVTLAAANRAKALKAQGRDIIDLTLGQPDFPTPKIIGQAAIEAIDNGKSFYTQ
AAT_BACSP       -LANRVKTLTPSTTLAITAKAKEMKDVIGLGAGEPDFNTPQNIMDAAIDSMQQGYTKYTP
Q9I015          --SNKLANVCYDIRGPVLKHAKRLHRILKLNIGNPAFEAPEEILQDVIRNLPTAQ-GYSD
Q9KQM1          -MSSKLDNVCYDIRGPVLKHAKRMHKILKLNIGNPAFDAPDEILVDVIRNLPTSQ-GYCD
Q9A8H2          ----RLPPYVFEEVNKIKALRAEGTDIIDFGMGNPDMPTPQHIVKLIETARDPKAGRYSA
O27916          -FASRVKDIQLSEIRKIFEVAD--EDTINLGIGEPDFSVPDHVREAVKDAVDEGLTHYTS
AAT_BACST       -LAKRVASLTPSATLAITEKAKELKDVIGLGAGEPDFNTPQHILDAAIKAMNEGHTKYTP
AAB1_RHIME      --ASRISSIGVSEILKIGARAAAMKPVIILGAGEPDFDTPEHVKQAASDAIHRGETKYTA
AATA_RHIME      ALSRVKPSATIAVSQKARELKAKGRDVIGLGAGEPDFDTPDNIKKAAIDAIDRGETKYTP
Q9HHD3          --AMSIEYAIRDVVLPARELEKQGIKIIKLNIGDPDFQPPEHMKKAYCEAIMEGHNYYGD
Q8Y525          -IAKKHQQMPVNILADIGTLAKTMPDILDLSIGDPDLITDESIINAAFEDVRAGHTKYTE
AAT_RICPR       -ISTRLNSIKPSPVKKTLELKKAGVNIIALGAGEPDFDTPDNIKEVAITSIKDGFTKYTN
Q8YIC8          ALSRVKPSATIAVSQKARELKAKGKDVIVLGAGEPDFDTPENIKQAAIAAINRGETKYTP
Q9HK41          -VVDDISFGAIVKIRQLLEMQRAGKKVYRLESGDPSFPIPEHVRLAIEDALKKNRTHYTD
AAT_RHILP       ALSRVKPSATIAVSQKARELKAKGRDVIGLGAGEPDFDTPDNIKKAAIDAINRGETKYTP
Q53951          -VSARVGAISESATLAVDAKAKALKPVIGFGAGEPDFPTPDYIVEAAVEAKNPKFHRYTP
Q9HKR7          ------ESATVSASNNVEKLRKQGLKIYNFGIGEPDFTTPEGIIDYAFEMAKQGKTHYTP
O25383          --SSKIQSLSESATIAISTLAKELKDILSFSAGEPDFDTPQAIKDAAIKALNDGFTKYTP
Q9CAP1          -VAKRLEKF-KTTIFTQMSILAVKHGAINLGQGFPNFDGPDFVKEAAIQAIKDGKNQYAR
AAT_THETH       -LSRRVQAMKPSATVAVNAKALELRDLVALTAGEPDFDTPEHVKEAARRALAQGKTKYAP
Q9RWP3          -LSARAQSLKPSATVAVTSRALELLDVISMSVGEPDFDTPPHVKAAGIAAIEEGKTKYTP
Q9UZ63          -IAERVILIKRSKIRELFERASKMENVISLGIGEPDFDTPKNIKEACKRALDEGWTHYTP
Q8U1F5          ALSDKLDLVNPSEIRKLFDLAAGMKDVISLGIGEPDFDTPAHIKEYAKEGLDKGLTHYGP
Q08415          -LHQSLTMTKRLQARRLDGIDQNLYDVVNLGQGFPDFSPPDFATQAFQQATSGNLNQYTR
AAT2_BACSU      --SDVIKTLPRQEFSLVFQKVKEMAHIINLGQGNPDLPTPPHIVEALREALNPSFHGYGP
Q9X8S5          PLLNRRLAEFGTTIFAEMSALAVRTGAINLGQGFPDTDGPEEVREAAVRALRDGRNQYPP
Q939K8          -FNQQTYKIEVSDIRRFDERVSVIEDMLKLTLGEPDFNTPEHVKLAGISAIENNDSHYTG
Q9RAT0          -FNPNLDKIEISLIRQFDQQVSSIPDIIKLTLGEPDFYTPEHVKQAGIAAIENNQSHYTG
Q9CJE0          -FNPNLDKIEISLIRQFDQQVSSIPDVIKLTLGEPDFYTPEHVKQAGIVAIENNQSHYTG
Q972A2          SISGESTLVYQDVARQVQKTKGI--RIINFGIGQPDLPTFARIREAAKKSLDEGFTGYTS
Q9XBE6          -MNRVLSALAPHSSKRISEYAQRHPGTVDLTVGLPAFGPPRAMLSSAPHVNARPEDQYAH
Q9HRM6          -FSTRVTSVEPSATLAVSTLASELVDVVDLSVGEPDFDTPESIVDAGKAAMDAGHTGYAP
Q92JE7          -ISTRLSSIKPSPVKKTLELKKAGVDIITLGAGEPDFDTPDNIKEGAIKAIKDGFTKYTN
Q9AA68          ----------PTTIFERMSGLARQYGAINLGQGFPDDQGPLPVREAAARALIEGSNQYPP
Q8U821          STADRLKNVSISATQRARELAAKGIKVVSLSSGEPDFPTPAHAIEAAHAAALNGETKYPP
Q92D16          SIRPELRDIEVSGIRTFNTRVTGIPDMIRLTLGEPDFPTPEHVKQAAITAIQENFTNYTP
O28650          -LARRVEELKPSGIRRFFDLVVGRDDVISLGVGEPDFPVPWRIREEMIYSLEKGYTSYTS
AAT_THEMA       -VSRRISEIPISKTMELDAKAKALEDVINLTAGEPDFPTPEPVVEEAVRFLQKGEVKYTD
Q8YP73          -LEKIPPYLFAEINRKREALIAKGVDIINIGVGDPDKPTPAHILQAMREAIDDASHNYPP
O53870          -MTVSRLRPYATTVFAEMSALATRIGAVNLGQGFPDEDGPPKMLQAAQDAIAGGVNQYPP
Q8TR00          SFSENVSRIDTSGIRKIFEAAGS--NAINLGLGQPDFDTPVHIKTAAIEAINEGFTGYTV
Q97I35          -FSKKAGQIAASITLEITAKADEMKNVIGFGAGQPDFNTPKNIRDAAIYAIENGYTKYTP
Q8XGH1          --SSKLENVCYDIRGPVLKEAKRLNKVLKLNIGNPAPFGFEAPDEILVDVIRNLPTGYCD
Q8UDD1          -LADILSRVKPSATQKARELKAKGRDVISLGAGEPDFDTPDNIKEAAIDAIKRGETKYTP
Q54188          -VSARVGAISESATLAVDAKAKALKPVIGFGAGEPDFPTPDYIVEAAVEACKNPKYHRYT
Q8Y606          PLSKRVKGVAPSPTLAITAKAKQMKDVIGLGAGEPDFNTPQNIIDAAIESMNKGFTKYTP
Q8TPT6          --SARLKRVEESATIRISNIATRMTDVINFSLGEPDFDTPKNICDAAAKAMYEGKTHYAP
Q98H83          AFDRLGEENAFAVLARATALAQQGRDIVNLGIGQPDFKTPQHIVEAAIKALRDGHHGYTP
O31665          -LKELPKQFFASLVQKVNRKLAEGHDVINLGQGNPDQPTPEHIVEEMKRAVDPENHKYSS
O28151          --ANRIEVVQPSATLRVSTMAKELKDVVDMSVGEPDFPTPDFIIEAAYKAMKEGKVFYTP
```

```
Q8Y8A4          SIRPELRDIQVSGIRTFNTRVTGIPDMIRLTLGEPDFPTPDHVKQAAISAIEENFTNYTP
Q92AB1          PLSKRVKGVAPSPTLAITAKAKQMKDVIGLGAGEPDFNTPQNIIDAAIESMNKGFTKYTP
O54170          ATPPASRIAELRRRSRRPALAPAPPGAVSLAMGEPDFPTPPTVVQAAVSALREGHTHYAD
Q9A830          ALRRIAPSATIAISAKARALKAAGRDVIALSAGEPDFDTPDNIKNAAIEAIKAGKTKYTD
Q8XRN7          AARDAIRDLRASRIREVANAGLGLPDVLPFWFGESDRVTPEFIRDTAAQALARGNTFYTH
Q8W360          -LSAASRRVAPSPIQQLSHLAQR-AGAVNLAEGFPDFPAPAHVKAAAAAAIAADLNQYRH
Q929C3          -IAKKHQQMPVNILADIGTLAKTMPDILDLSIGDPDLITDASIINAAFEDVRAGHTKYTE
O86587          -LSEVCYEIRGPVIEHANALEEAGHSVLRLNTGNPALFGFEAPEEIVQDMIRPRAHGYTD
O59044          ----SIEYAIRDVVLPARELEKKGIKVIRLNIGDPDFQPPEHMKEAYCRAIQEGHNYYGD
Q943I5          -ISPTVSALRPSKTMAITDQATALRPVIGLAAGEPDFDTPHVIAEAGMNAIKDGYTRYTP
Q97TA8          -FNKQLDKIQVSLIRQFDQAISEIPGVLRLTLGEPDFTTPDHVKEAGKRAIDQNQSYYTG
AAT_STRVG       -MSARIGAISESATLAVDAKAKALKPVIGFGAGEPDFPTPDYIVEAAVEARNPKYHRYTP
Q8RQG1          ------------------------------------------------------------
Q8XQJ3          ELNRHLAAAQPSATYRVIDRVAARAEVISLSAGEPDFDTPAHVREAGIEAIRAGMTRYTQ
Q8TQ40          -INALPPYLFAAIDEAKDEMIAKGVDVIDLGVGDPDLPTHPHIVEAMREAVCDPKTQYPS
AAT_RICCN       -ISTRLSSIKPSPVKKTLELKKAGVDIITLGAGEPDFDTPDNIKEGAIKAIKDGFTKYTN
Q9RWJ7          ---------AQTSIFTHMSLLAARHGAVNLGQGFPSNPPPAFLLDAARRAVGT-VDQYTP
Q97GI7          -ISKGANSIQISGIRKFYNKVIKVQGAISLTLGQPDFPVPDKVKRAMVRAIEDNKTVYTS
O95335          ------------------------------------------------------------
Q98I67          -LADTLSRVKPSATIAVTQKARELKDIIGLGAGEPDFDTPDNIKNAAIEAIRRGETKYPP
AATC_RHIME      ----RLPPYVFEQVNRLKASAAAGADIIDLGMGNPDLPTPQSIVDKLCEVVQDPRTRYSS
AAT_STRGR       -VSARIGAISESATLAVDAKAKALKPVIGFGAGEPDFPTPDYIVDAAVEACRNPKYHRYT
Q8YA73          -LQNLPDQFFSSLVEKVGKKVAEGHDVINLGQGNPDQPTPKHIVEAMKTASEKPLHKYSL
Q8ZI88          ----------------------------------PDPVLPEHISQAVIKSMENGSSHYTM
O67864          ----RLPQYVFSLVNELYKLRREGEDVVDLGMGNPNMPPAKHIIDKLCEAQKPNVHGYSA
Q93RH7          -LKKLPKQFFADLVTKVNAKIAQGADVINLGQGNPDRPTYDFIKALQDSAAKPASHKYSL
Q8R7H1          -ISDVVKKIPPSGIRKFFDLVTNSKDIISLGVGEPDFVTPWEIRKEGIETLCRGNTTYTS
Q9A0S0          -VLEMKESVTLAAGARAKALKAQGRDVLNLTLGEPDFFTPKHIQDKAIESIQNGTSFYTN
AAT_SULSO       -FNGNMSQVTGETTLLYKEIARNVEKIIDFGIGQPDLPTFKRIRDAAKEALDQGFTFYTS
Q8ZW57          ----------------------------RLDIGDPDLPPPPELLEALGRV---GDMRYGP
Q982E0          -------------------------EIVDLTAGEIWSELARMIRQGAIEAINKGVNRYTD
Q982E3          -LSQRMSLLSAQEPAELAQLAAAARQIIDLAAGEIIIETPLSVREGAIAAINAGTNRYTD
O66630          EFSDRLKVLPAELDRKKQEKIEQGVDVIDLGVGDPDMPTPKPIVEAAKKALENPEHKYPS
AAT2_METJA      -LSKRLLNFESFEVMDILALAQKLKKVIHLEIGEPDFNTPKPIVDEGIKSLKEGKTHYTD
O29838          -LADRLEKIPPYLFAEIDAMKRKKVKVIDFGVGDPDLPTPEHIVEALKNAAEKVERKYPS
Q9HQK2          --SERAAAVTPFAAMDVLERAADRADVIHMEVGEPDFAPPAAATEAAVDALRAGDDDYTT
AAT1_BACSU      -LAKRVSALTPSTTLKAKELKAAGHDVIGLGAGEPDFNTPQHIIDAAVRSMNEGHTKYTP
Q9YE99          --NRRVDLITGSPTRKIDAVRERLRDVILLSTGQPGFLPPTFLRERLAQALLKRLYSYTP
Q9R6Q3          ------ESVTLAAANRAKALKAQGRDIIDLTLGQPDFPTPKKIGQAAIEAINNGQSFYTQ
Q97AE8          -IVDEISFGAIVKIRQLLEMQRQGKKVYRLESGDPSFSLPPHVKEAIKQAIENNKTHYTD
O30304          --AARLSQISTSMIRRMFEIVERAKEIISLTIGEPDFDTPQEVIERACRAMNAGFTHYTS
Q92QJ6          ----QMSKRSAVEPFHAMDVLAEAHPVISMAVGQPAHPAPKAALDAARRALDHGRLGYTD
AAT_PYRHO       -IAERVLLIKRSKIRELFERASKMEDVISLGIGEPDFDTPKNIKEAAKRALDEGWTHYTP
Q98KW9          ---DGWEVHFAAWTRK-----EAGEDIIMLSVGDHDFDTPSQTIEACVTAVRGSNHHYTP
Q8VS38          -LAKRMSLIKPSPTIAVTDKANRLKKICILAAGEPDFDTPDHIKKVAIQAIDEGKTKYTA
Q8XMH8          -ILDNVKNMPPSGIRKYFDLINEMEDVISLGVGEPDFVTPWNVREAGIYSLEQGHTHYSS
AAT_PYRAB       AMSDRLDLVNPSEIRKLFDIAAGMKDVISLGIGEPDFDTPQHIKEYAKEALDMGLTHYGP
Q8VS39          -----MSLIKLSPTIAVTDKVNRLKEICVLAAGEPDFDTPGHIKKAAIQAINEGKTKYTA
Q9KE01          ------EQFFAKLVEQVQEVKKDHDDIINLGQGSPDLPTPEHIVEKLQEAAENPLHRYAP
O05237          -LSDYVQQIKPSGIRKFFDLAATMEGVISLGVGEPDFVTAWNVREASILSLEQGYTSYTA
AAT_AQUAE       -LASRVSHLKPSPTLTITAKAKELRDVIGFGAGEPDFDTPDFIKEACIRALREGKTKYAP
Q8REF4          -ISDRVKNMKYSAVRKLAPLAAEAEKVYRLNIGQPNIETPKLFFEGLKNIPDHVI-RYAD
Q9PPF7          -LTKRSQVLEESITLAITALANELKDIISFSAGEPDFDTPQTIKNAAISAIEKGCGKYTA
Q8Y1I0          -LANIRAFHVMELAKQARELELAGRSIIHMGIGEPDFTAAEPVVRAAEAAMRRGVTQYTG
```

```
Q8RR70          -LAARVESVSPSMTLIIDAKAKAMKDVCSFSAGEPDFNTPKHIVEAAKAALEQGKTRYGP
Q9HRX4          EFSDRVAQVSISGIRAVFEAAG--EDAINLGLGQPDFPTPDHARQAAVDAIESGAAGYTS
Q97M25          -INPLVKKIELSPIREISDAALKYSDAINLTVGQPDFTTPEHIKLSAKKAIDNNHTSYTA
Q9P9M8          ----SVEYAIRDVVLPARELEKKGIKVIRLNIGDPDFQPPEHMKEAYCKAIKEGHNYYGD
AAT_THEAQ       -LSQRVKSMKPSATVAVNARALELRDLVALTAGEPDFDTPEHVKEAGRRALAQGKTKYAP
Q8XXV2          ----RLPKVGTTIFTVMSALAAE-KQAVNLGQGFPDFDCDPRIVDAVSDAMRAGFNQYPP
Q8TS80          -VAEAVKSIPPSGIRRFFDLVSGLEDIISLGVGEPDFITPWHIREMCIHSLEKGQTSYTS
Q8VXZ8          -VAKRLEKF-KTTIFTQMSILAVKHGAINLGQGFPNFDGPDFVKEAAIQAIKDGKNQYAR
Q8XRB5          -IAARVRRIKPSPSTSAADRANELKSIVNLVVGEPDFDTPQHIRQAAVQAIERGATRYTL


1gdeA           NIGLLELREAIAEKLKKQNGIEADPKTEIMVLLGANQAFLMGLSAFLKDGEEVLIPTPAF
Q8Y0E8          ISGLARLREAVARKFRDENGIE-CTAAETLVGCGGKQVIYQAFVATIDPGDEVLIPAPYW
Q9BVY5          ----------------------------------------------------------ATPVF
Q8RAK7          ASGILELKKAICEKLKRENGLFYEP-EQIVVSNGAKHSIYNALSAILNPGDEVIIPVPYW
AAB2_RHIME      LDGTPELKKAIREKFQRENGL-AYELDEITVATGAKQILFNAMMASLDPGDEVVIPTPYW
Q55453          VAGIPSLISALTEKLRRDNDINLSSDQAVVVTAGANMGFLNAVLAITEVGDEIILNTPYY
AATB_RHIME      LDGTPELKKAIREKFQRENGL-AYELDEITVATGAKQILFNAMMASLDPGDEVVIPTPYW
Q8YTF2          FEGTASFRRAITNWYNRRYGVVLDPDSEALPLLGSKEGLSHLAIAYVNPGDVVLVPSPAY
Q9KC79          SGGLPKLKEAIIEKFKQDQGLTYTAK-EIFVGTGAKHVLYTLFQALLNEGDEVIIPSPYW
Q92S71          QRGIPPLREALVRYYQRRFQKALSP-ENFYVTGSGMQAIKLAIEAVGSPGDEVVLLTPAW
Q93703          AVGALAAREAIVERYSSADNVF--TADDVVLASGCSHALQMAIEAVANAGENILVPHPGF
Q8YMS6          AAGEPKLREAIARKLQKDNHLDYKP-ENVIVTNGGKHSLYNLIVALIDPGDEVIIPAPYW
Q9HV76          ARGLPALREAIAKFYGERYGVDLDP-QRVLVTPGGSGALLLASSLLVDPGRHWLLADPGY
Q8U3E6          NAGIPELREAIAEYYKKFYGIDVEV-DNVLVTAGAYEATYLAFETMLEQGDEVIIPDPAF
Q9HUI9          VRGKRALRQRIAERHRRRSGQAVDA-EQVVVLAGAQCALYAVVQCLLNPGDEVIVAEPMY
Q8ZVJ5          SQGIYEVRQAISEDLRRLGGLEVPP-EQIVLTAGGQAAMFSTLATLIEPGDEVVVTDPTY
AAT_METEX       SKGIAGLRRAQAGYYQRRFGVSLNPDT---------------------------------
Q9ZLG5          VAGIPELLKAIAFKLKKENNLDYEP-SEILVSNGAKQSLFNAIQALIGEGDEVVIPVPFW
AAT1_METJA      NNGIPELREEISNKLKDDYNLDVD-KDNIIVTCGASEALMLSIMTLIDRGDEVLIPNPSF
Q8XJ54          NAGIVELREEISSLLKNNFDID-FSKDEIIITVGGSEGLYAAMTALLNPGEKVLVPSIAY
Q8UFR3          ALGLRELREAIAGHYRLRHQVAIDP-ARIAVTTGSSAAFNLAFLGLFDAGDHVAIARPGY
Q9W6U2          AFGHVPLVKSLAKFFSRVIGHEIDPLEDILVTVGAYQALFSAFQALIYEGDEVIIVEPFF
Q9K7L1          NAGIIELREAISRYLYERFHIGYDPESEILVTVGASEAIDIGMRAIIDEGDEVIVVEPSF
O66737          ALGLWELRERISEFYRKKYSVEVSP-ERVIVTTGTSGAFLVAYAVTLNAGEKIILPDPSY
Q9Y9P0          TAGIPELREAIAWYLNSRYGADVSP-EEVIATTGAKTAIFLGMALYLRPGDEVIIPDPSY
Q99V44          NKGLLETREAISQYFKNRYHFSYDP-EEIIVTNGASEAIDTTLRSIIEPGDEIIIPGPIY
Q8RA61          SSGIPELKKAICEKLLKDNGLSYTP-EQIVVSNGAKHSIYNALSAILNPGDEVIIPVPYW
PATA_BACSU      NAGYLELRQAVQLYMKKKADFNYDAESEIIITTGA-QAIDAAFRTILSPGDEVIMPGPIY
Q98NB8          TLGLAGLRKAIAEHYADHYRLEVEP-ARIAVTTGSSAAFNLAFLAMFDPGDRVAIAAPGY
Q97PQ9          TSGLPELKAAVNSYFERFYGYSV-ASNQVTVAAGAKYSLYTFFMAVVNPGDEVIIPTPYW
Q9KAU1          NQGLPEARQKVAEHMNGRFNTNITAQTVTMT-SGAAGALNVALKSIMNPGDEVIIFTPYF
Q9V0G5          SEGLMELREAIVEREKKKNGVNITP-DDVRVTAAVTEALQMIFGALLDPGDEILIPGPSY
AAT_PYRKO       TGGIRELRERIAE-------FEGVSADEVIVAPGAK----ILIAAEIASAKKVAVVSPRW
Q9L0L5          AGGLPELKAAIAAKTLRDSGYEVDP-SQILVTNGGKQAIYEAFAAILDPGDEVIVPAPYW
Q979X6          SNGIHELREKVSEKLKNRNNINASP--DEVLITPTKFGINLAMMVILNPGDEVLIPEPYY
Q8RCV4          NAGLIELRREIADYLKRRYSLYYDPEKEILVTIGATEAIYVTLSTLAEEGDEVLIPEPSF
AAT_SYNY3       AAGEPALRQAIAKKLREKNNLPYEA-ANILVTNGGKHSLFNLMLAMIEQGDEVIIPAPYW
Q8XJT3          TSGIVELKEAIARKLHDDNGLNYGTK-NIIISTGAKQSLANVFMAILNPGDEVIIPVPYW
Q97KB8          SQGIDTLIDSFIRSLSDSNIYF--EKDEIMITHGGSEAIQFAIMATCDPGDEILSPEPFY
Q8ZDK4          SKGLFSARKAIMQHYQARN-IRDLTVEDIYIGNGVSELIVQSMQALLNLGDEMLVPAPDY
O59096          NIGLLELREAIAEKLKKQNGIEADPKTEIMVLLGANQAFLMGLSAFLKDGEEVLIPTPAF
Q98AR6          TVGMVELREALARKISLDTG-QIWKAEEVAVTSGAKQALFNAAMVLLNPGDEVIIPAPYW
Q8TT07          SQGLPALREAIAESYYRKFGVDLDP-NQVIVTSGTSPGLLMVFMALLEKRDEVIMSNPHY
```

```
Q9R096          AFGYPPLTNVLASFFGKLLGQEMDPLTNVLVTVGAYGALFTAFQALVDEGDEVIIMEPAF
Q9RNK6          VDGTAELKEAIVGKFRRDNHLE-YRTDQISVGSGGKHVLFNALTATIDQGDEVIIPAPYW
Q9SIE1          NAGITELREAICRKLKEENGLSYAP-DQILVSNGAKQSLLQAVLAVCSPGDEVIIPAPYW
YD91_METJA      N-GIQELKDAVPPYMEKVYGVKDDPVNEVIHSIGSKPALAYITSAFINPGDVCLMTVPGY
Q9CPI6          SKGLYSARKAIVQYYQS-KGIHGATVNDVYIGNGVSELITMSLQALLNDGDEVLIPMPDY
Q8U097          TTGINELREKIAE-------VEGVSKEEVIVGPGAK----ILIAAEIAMANKIGVIAPYW
Q9CEK7          AGGLPELKNAVQHYWTRFYNYEIQP-NEILITAGAKFALYAYFMATVDPDDEVIIPAPYW
AAT_BACSP       SGGLPALKQAIIEKFKRDNQLEYKP-NEIIVGVGAKHVLYTLFQVILNEGDEVIIPIPYW
Q9I015          SKGLFSARKAVMQYYQ-QKQVEGVGIEDIYLGNGVSELIVMSMQALLNNGDEVLIPAPDY
Q9KQM1          SKGIYSARKAVVQYYQK-KGIRSLDVEDVYIGNGASELIVMAMQALLNNGDEMLVPAPDY
Q9A8H2          SKGVPGLRKAMANYYGRRFGVKLNPDTEVIATLGSKEGFANLAQALTGPGDVIICPNPAY
O27916          NMGMEELREAIADKLKSENRVHAEP-ESIIVTVGASEAIFMCTQALLDIGDHALIPDPGF
AAT_BACST       SGGLPALKEEIIKKFARDQGLDYEP-AEVIVCVGAKHALYTLFQVLLDEGDEVIIPTPYW
AAB1_RHIME      LDGTPELKKAIREKFQRENGL-AYELDEITVATGAKQILFNAMMASLDPGDEVIIPTPYW
AATA_RHIME      VSGIPELREAIAKKFKRENNLD-YTAAQTIVGTGGKQILFNAFMATLNPGDEVVIPAPYW
Q9HHD3          SEGDRELREAIVEREKKKNGVDITP-EDVQVTAAVTEALQFIFGALIDGGEEILIPGPSY
Q8Y525          SGGDVELIDAIRGYFSRNYDLS-FERSQIRATVGALHGMYLTLQTILDDGDEVIIHEPYF
AAT_RICPR       VDGIPLLKQAIKNKFKRENNID-YELDEIIVSTGGKQVIYNLFMASLDKGDEVIIPVPYW
Q8YIC8          VSGIPQLRQAIVSKFKRENGLDYKP-EQTIVGTGGKQILFNAFMATLNPGDEVIIAAPYW
Q9HK41          STGIPELRKAIAEKLVKKNGIRSATPENVLVSNGGMNALYITFRSLIAPGEKVIIPDPMW
AAT_RHILP       VSGIPELRKAIAAKFKRENGLD-YSWEQTIVGTGGKQILFNAFMATLNPGDEVSIPAPYW
Q53951          AGGLPELKAAIAAKTLRDSGYEVDP-SQILVTNGGKQAIYEAFAAIL-------------
Q9HKR7          SAGIMELREKIASKLKTRNRIDANA--ENVLVTPTKFGINLAMMVILNPGDEVIIPDPYY
O25383          VAGIPELLKAIAFKLKKENNLDYEP-NEILVSNGAKQSLFNAIQALIEEGDEVIIPVPFW
Q9CAP1          GYGIPQLNSAIAARFREDTGLVVDPEKEVTVTSGCTEAIAAAMLGLINPGDEVILFAPFY
AAT_THETH       PAGIPELREALAEKFRRENGLSVTP-EETIVTVGGKQALFNLFQAILDPGDEVIVLSPYW
Q9RWP3          VSGIPELREAISAKFRRENGLDYAP-NAVTVTSGGKQALFNAFFALLNPGDEVLIPAPHW
Q9UZ63          NAGIQQLREAVVEYYKKFYDVDIDV-ENVIITAGAYEGTYLAFESLLESGDEVLIPDPAF
Q8U1F5          NVGLPELREAIAEKLKKQNGIEADPNSEIMVLVGANQAFLMSLATFLKDGEEVLIPSPMF
Q08415          AFGYPPLTNVLASFFGKLLGQEMDPLTNVLVTVGAYGALFTRFQALVDEGDEVIIMEPAF
AAT2_BACSU      FRGYPFLKEAIAAFYKREYGVTINPETEVALFGGGKAGLYVLTQCLLNPGDIALVPNPGY
Q9X8S5          GPGVPELRAAVAGHQQRRYGLSYDPDTEVLVTAGATEAIAAALLALLEPGDEVVALEPYY
Q939K8          MAGDLELRKAVATFMQKKYQVSFAPENEILVTVGATEALSASLLAVLNPGDKIIVPTPIY
Q9RAT0          MAGLLELRQAASEFLLKKYGLSYAAEDEILVTVGVTEAISSVLLSILVAGDEVLIPAPAY
Q9CJE0          MAGLLELRQAASEFMNKKYGLSYAAEDEILVTVGVTEAISSVLLSILVAGDEVLIPAPAY
Q972A2          AYGIDELRQKIAEHL--SSKYESVRKEEVIVTPGAKTALYLAFLLYINPGDEVIIFDPSF
Q9XBE6          SRGAIELRAAIAHVYKSEQGVDLDPDTQILVTNGAAGALWIAVLTLTEPGDEVLLADPGY
Q9HRM6          SNGVPELRDAIAEKL--QGDGLDYEAGNVIVTPGAKQALYETFQAVVDEGDEVALLDPAW
Q92JE7          VEGMPLLKQAIKDKFKRENNID-YELDEIIVSTGGKQVIYNLFMASLDQGDEVIIPAPYW
Q9AA68          MRGLPELRAAVAGHYGRTQDLTLDPDTEIVVTSGATEALAAAFTSLISPGDEVVLFQPLY
Q8U821          MDGTVAMKAAISRKFKRDNNLTYDA-SQIVVSAGGKQVIFNAMLATCNPGDEVVIPAPSW
Q92D16          NAGMPELLEAASSYFEEKYDLT-YSNKEIIVTVGATEAISVALQTILEPGDEVILPDPIY
O28650          NLGLPELREGIAEYYTR-FGVKALP-EQVMVTSGVSEGVDIAIRALIEPGDAALIPEPCY
AAT_THEMA       PRGIYELREGIAKRIGERYKKDISP-DQVVVTNGAKQALFNAFMALLDPGDEVIVFSPVW
Q8YP73          YEGTQEFREAAVEWMERRFGVMDNPNTEVVSSIGSKEAIHNTFLAFVEAGDYTLIPDPGY
O53870          GPGSAPLRRAIAAQRRRHFGVDYDPETEVLVTVGATEAIAAAVLGLVEPGSEVLLIEPFY
Q8TR00          GPGIPELREALSQKFLEENGFSVSP-QEIIVTSGASEALTIALAALLNVGDEVLISNPGF
Q97I35          VSGIKELKMAICDKFKRDNNLN-YSLSNIIVSTGAKQCLSDTFSALLNPGDEVILSAPYW
Q8XGH1          SKGLYSARKAIMQHYQAR-GMRDVTVEDIYIGNGVSELIVQAMQALLNSGDEMLVPAPDY
Q8UDD1          VSGIPELRKAIADKFKRENGLDYKP-EQTIVGTGGKQILFNAFMATLNPGDEVVIPAPYW
Q54188          RRRLPELKSAIAAKTVRDSGWEPDV-SQILVTNGGKQAIYEAFAAILDPGDEVIVP----
Q8Y606          SSGIIELKQAIVDKLKKDQFLN-YETNQIFVGTGAKHVLYSAFQTILDPGDEVIIPVPYW
Q8TPT6          SAGIPELRAAIAEKLKTENHLEVTEK-DVLVTPGAKQAIFEIMMGALDDGDRALLFDPAW
Q98H83          ANGLLATREAVVRRTLTTTGVEVSP-EAVMILPGGKPTMFAAILMFGEPGAEILYPDPGF
```

```
O31665          FRGSYRLKSAAAAFYKREYGIDLDPETEVAVLFGGKAGLVELPQCLLNPGDTILVPDPGY
O28151          TKGVPELIDAIVEKLRNENGIDVGA-ENIIVTPGAKYAIFEAMMCLLQEGDEVILLDPSW
Q8Y8A4          NAGMPELLEAASIYFHEKYDLS-YSNKEIIVTVGATEAISVALQTILEPGDEVILPDPIY
Q92AB1          SSGILELKQAIVDKLKKDQSLTYEA-NQIFVGTGAKHVLYSAFQTILDPGDEVIIPVPYW
O54170          QRGLRELRAALAARLPERPGGAWDA-DDVLVTHGATAALAAVVLATVGPGDRVVVPEPAY
Q9A830          PDGMPELKAAICAKFKRENGLEYKP-SQIHVAPGGKPVIYNALVATLNPGDEVIIPAPYW
Q8XRN7          NLGIAPLRSALADYVSRLHGRTA--IDHVAVTSAGVNALMLAAQLVVGPGDRVVTVTPLW
Q8W360          ---VQGICDALAETMKRDHGLRVDPLTDFAVCCGQSEAFAAAIFAIIDQGDEVLLFDPAF
Q929C3          SGGDVELIDAIRGYFSRNYDLS-FERSQIRATVGALHGMYLTLQTILDDGDEVIIHEPYF
O86587          SRGILSARRAVAQRYQAL-GLEVDV-DDVFLGNGVSELISMAVQALLEDGDEVLIPAPDF
O59044          SEGLIELREAIVKREKEKNGVDITP-DDVRVTAAVTEALQLIFGALLDPGDEILIPGPSY
Q943I5          NAGTLELRKAICNKLQEENGISYSP----------------------DQVLVLIPAPYW
Q97TA8          MSGLLTLRQAASDFVKEKYQLDYAPENEILVTIGATEALSATLTAILEEGDKVLLPAPAY
AAT_STRVG       AGGLPELKAAIAAKTLRDSGYEVEA-SQVLVTNGGKQAIYEAFAAILDPGDEVIVPAPYW
Q8RQG1          ------------------------------------------------------------
Q8XQJ3          VAGLRALREAVADKFRGENGL-AVGWQDTIVCSGGKQVIYNALAATLNEGDEVVIPAPYW
Q8TQ40          YAGMPEFREAAAEWCKKYKGIELDPATEVLSLIGSKEAVAHIPLAFVNPGDVVLYTDPGY
AAT_RICCN       VEGMPLLKQAIKDKFKRENNID-YELDEIIVSTGGKQVIYNLFMASLDQGDEVIIPAPYW
Q9RWJ7          PLGLPALREALGADLNV-------DPADVIVTSGATEGLLTLALSLLGQPAELVVFEPVY
Q97GI7          NAGIDELRNEISKYLKRFNINY--SKDEICITAGGTEGILDIFQALLNKGDKVLVPDPSF
O95335          ----------------------------------------------------ATPVF
Q98I67          VSGIVPLREAIAKKFKRENNLDYKP-EQTIVGTGGKQILFNAFMATLNPGDEVIIPRPYW
AATC_RHIME      SKGIPGLRRAQAAYYARRFGVKLNPETQVVATLGSKEGFANMAQAITAPGDVVLCPNPTY
AAT_STRGR       PQRAPELKAAIAEKTLRDSGYEVDA-GQILVTNGGKQAIYEAFAAILDPGDEVIVPAPYW
Q8YA73          FRGKQELKQAAADFYAREYNVTIDPNTEVAILFGTKTGLVELPMCLMDPGDTMLLPDPGY
Q8ZI88          PIGNPELKEKIALKLQRYNNLTVEAQRNILITPGSDSGLLFAMMPFINNDDEVLIHSPSY
O67864          SRGIPRLRKAICNFYEERYGVKLDPEREAILTIGAKEGYSHLMLAMISPGDTVIVPNPTY
Q93RH7          FRGNPPFKEAAADFYKANYQVDLDSQTEICVLGGSKIGLVELPWALMNPGELLLLPDPGY
Q8R7H1          NLGLLELRIAISYFLKTHYDLNYDPEKEIMVTIGASEAIDLALRALLNDGDEVLIPEPSY
Q9A0S0          ASGLPELKAAIATYLKNQYGYHLSP-DQIVAGTGAKFILYAFFMAVLNPGDQVLIPTPYW
AAT_SULSO       AFGIDELREKIAQYLNTRYGTDV-KKEEVIVTPGAKPALFLVFILYINPSDEVILPDPSF
Q8ZW57          PEGLQAFREAVASIFGVDPG-------EVVAVAGGRHGLAALMWIFRK--RRLLTTRPYY
Q982E0          PVGTTELREALARKVSLETR-QIWKAEEIAVTCGAKQALFNAAMVLLDPGDEVIIPVPYW
Q982E3          AIGLTLLRKAVAEKLAAQTHVGWN-LEDIVITAGAKQALLNAALAVLDPGDEVIIIRPSW
O66630          YVGKYEFRKAVADWYKRRFDVDLDPNTEVITLIGSKEGIAHFPLAFVNPGDIVLCPDPAY
AAT2_METJA      SRGILELREKISELYKDKYKADIIP-DNIIITGGSSLGLFFALSSIIDDGDEVLIQNPCY
O29838          YEGMLSFRESVARFYRRRKGVNLDPESEVISLIGSKEGIAHLPLAFVNDGDYVLVPEPGY
Q9HQK2          SRGRRSLRDAISGYYAAEYGVSV-PAERIVVTPGSSPALLTVLLATVDPGSAVVLSDPHY
AAT1_BACSU      SGGLAELKNSIAEKFKRDQNIEYKP-SQIIVCTGAKHALYTLFQVILDEEDEVIIPTPYW
Q9YE99          TPGYADVREAIAEDLAALGGPRMEP-DDILVTAGGQEAMFATLSTILEPGDKVILMDPTY
Q9R6Q3          AGGLPELKKAVQHYWTRFYAYEIQT-NEILITAGAKFALYAYFMATVDPLDEVIIPAPYW
Q97AE8          STGIPELRKAIAEKLVRKNKIKDATPENVIVSNGGMNALYVTFRSLLSPGDEVIIPDPMW
O30304          NFGLEELRSAIAERYGV-------DSSNVMVTAGGSEALLNASLAFIEEGSKVVIPSPNF
Q92QJ6          ALGTHSLKRAIAAHYHSRHGITLDP-QRIAVTTGSSAGFNLAFLALFDPGDRVAIARPGY
AAT_PYRHO       NAGIPELREAVVEYYKKFYGIDIEV-ENVIITAGAYEGTYLAFESLLERGDEVIIPDPAF
Q98KW9          LPGLPRLRQAMAAASSACTGIETTP-DQVIATPGGQAALYAAVQAVLDQGDHAIVVAPYY
Q8VS38          VDGTRELKEAIISKLKRDNNLE-YALSQVCVGAGAKQVLFNLFMATVNPGDEAIIPAPYW
Q8XMH8          NAGFIELREEISKYLNRRFSLRYNPKDEILVTVGGSEGIDLALRALVGPGDEVIIPEPSF
AAT_PYRAB       NIGLPELREAIAEKLKKQNNIEADPNKEIMVLVGANQAFLMGLSAFLKDGEEVLIPTPAF
Q8VS39          VDGTRELKGAIIDKLKRDNDLEYMP-SQICVSAGAKQVLFNLLMATINPEDEAVIPAPCW
Q9KE01          FSGYPFLKEAVSKYYEREYGVSVDPKTEVAVLGGAKTGLVEVSQCFLNPGDMALVPDPGY
O05237          NAGLYSLREEISRYLSNRFDLSYSPDNELIVTVGASQALDIAIRAIVNPGEEVIIPEPCF
AAT_AQUAE       SAGIPELREAIAEKLLKENKVEYKP-SEIVVSAGAKMVLFLIFMAILDEGDEVLLPSPYW
Q8REF4          SRGISILLEQVIEVYARDGHIL--KKEDIIVTEGGSEALTFAMLAICNPNDEVLIPEPFY
```

```
Q9PPF7          VAGIPEVLKAIQTKFKKDNNLD-YETNEIITNVGAKHSLFECIECLVEKDDEVIIPSPYW
Q8Y1I0          ALGIRPLREAIARYYHTVYGLDIAP-ERIIVTAGASAALLLACAVLVEIGGEVLMPDPSY
Q8RR70          AAGEPRLREAIAQKLQRDNGLC-YGADNILVTNGGKQSIFNLMLAMIEPGDEVIIPAPFW
Q9HRX4          NRGTAALVDAIVEKHARDQGVDVAP-AGVIATAGGSEALHLAMEAHVDPGDEVLFPDPGF
Q97M25          NSGICELRKAASNFINKKYNLNYNSDKEIIVTNGATEAIDISLRTILEKDDEVLLPAPIY
Q9P9M8          SEGLPELRKAIVEREKRKNGVDITP-DDVRVTAAVTEALQLIFGALLDPGDEILVPGPSY
AAT_THEAQ       PAGIPELREAVAEKFRRENGLEVTP-EETIVTVGGKQALFNLFQAILDPGDEVIVLAPYW
Q8XXV2          MTGVPALRQAIAAKIATLYGHAYDAEREITVTAGATQALLTAVLCCVHPGDEVIVFEPTY
Q8TS80          NYGLPELRDELARTYYKRYGLDYDPASEILVTTGVSEALDIAVRAVVNPGEEVIVVQPSY
Q8VXZ8          GYGIPQLNSAIAARFREDTGLVVDPEKEVTVTSGCTEAIAAAMLGLINPGDEVILFAPFY
Q8XRB5          MAGTVELRQAIVDKMARENDLH-YAMNEIIATNGAKSAIYSALAITLEAADEVLIPAPYW


1gdeA           VSYAPAVILAGGKPVEVPTYEEDEFRLNVDELKKYVTDKTRALIINSPCNPTGAVLTKKD
Q8Y0E8          SSYADIVTLCGGIVKPLPTTPESGYALQPQTLAAGISARTKWLVLNAPSNPSGTAYTAAQ
Q9BVY5          IPLR-------SKPVYGKRWSSSDWTLDPQELESKFNSKTKAIILNTPHNPLGKVYNREE
Q8RAK7          LSYPEMVRLAYGKPVFVQTKEENNFKITAEELTAAINPKTKALILNSPNNPTGAVYTRKE
AAB2_RHIME      TSYSDIVQICEGKPILIACDASSGFRLTAQKLEAAITPRTRWVLLNSPSNPSGAAYSAAD
Q55453          FNHEMAVRIAGCQPVLVPT--DDQYQLQLDLIAQAIAPRTRAVVTISPNNPTGAIYPEAD
AATB_RHIME      TSYSDIVQICEGKPILIACDASSGFRLTAQKLEAAITPRTRWVLLNSPSNPSGAAYSAAD
Q8YTF2          PAHFRGPVIAGGTVHSLILKPENDWLIDLTAIPEEVARKAKILYFNYPSNPTGATAPREF
Q9KC79          VSYPEQVKLAGGEPVFVEGKESNDFKLTPAQLEPVLTDRTKAIIINSPSNPTGSLYTQEE
Q92S71          PNFAAAADLSGVRPVAVPLFEGGKWRLDPERLQAAIGERTRALFINTPSNPTGWTATHDD
Q93703          PLYSPHNIVDKPYKIDMTG---EDVRIDLSYMATIIDDNTKAIIVNNPGNPTGGVFTKEH
Q8YMS6          LSYPEMVTLVGGKSVIVPTDASTGYKITPEQLRKAITPKTKLFVLNSPSNPTGMVYTPEE
Q9HV76          PCNRHFLRLVEGAAQLVPVGPDSRYQLTPDLIERHWDSDSVGALVATPANPTGTLLDRDE
Q8U3E6          VCYVEDAKLAEAKPIRLPLREENDFKPDIDELLERITKRTRMIVINYPNNPTGAVLDKET
Q9HUI9          VTYEAVFGACGARVVPVPVRSENGFRVQAEEVAALITPRTRAMALNSPHNPSGASLPRAT
Q8ZVJ5          FGYKPLLEYFGAVVKPVRTRLEDGFQPNPEALKDSVSRKTKALILVSPDNPTGRALKEEA
AAT_METEX       ------------------------------------------------------------
Q9ZLG5          VTYPELVKYSGGVSQFIQTDEKSHFKITPKQLKDALSPKTKMLILTTPSNPTGMLYSKAE
AAT1_METJA      VSYFSLTEFAEGKIKNIDL--DENFNIDLEKVKESITKKTKLIIFNSPSNPTGKVYDKET
Q8XJ54          PAYESISKIIGCEVINYDL--NEDFSVNIESLKEGIKEGGKLLVLSYPCNPTGALLSKKS
Q8UFR3          PAYRNILKALGLNVVEVPVTAETGYTLTPASLERAETKKLKGVLLASPANPTGTVTGREA
Q9W6U2          DCYQPMVKMAGGQPVYIPLRSSGDWVLSPEXLAGKFTPRTKALVINTPNNPLGKVYKTEE
Q9K7L1          VSYAPLVTMAGGVPVPVGTHIDTDFQVTPAQIEAAITPRTKAIILCFPNNPTGSIMGKEE
O66737          PCYKNFAYLLDAQPVFVNVDKETNYEVRKEMIE---DIDAKALHISSPQNPTGTLYSPET
Q9Y9P0          YAYAQVAKLFGARPVYVPMKFEPGFRFDIEGIERAVSEKTRMIVVNNPHNPTGSVFPPDQ
Q99V44          AGYIPLIEVLGGKPIYIDT-TATQFKITPDALESHISPKTKAVLLNYPTNPTGVVLKRNE
Q8RA61          LSYPEMVRLAYGKPVFVHTKEENDFKITAEELENAITPKTKAIILNSPNNPSGAVYTKEE
PATA_BACSU      PGYEPIINLCGAKPVIVDT-TSHGFKLTARLIEDALTPNTKCVVLPYPSNPTGVTLSEEE
Q98NB8          PAYR-NIMAALGIDVVEIELGDAAYLHAGHLETAHREKPLKGVLFASPANPTGAVIPADE
Q97PQ9          VSYGDQVKMAEGVPVFVSAKEDNHFKVTVEQLEAARTDKTKVLVLNSPSNPTGMIYTREE
Q9KAU1          AEYKFYVGNANGVAVYCPL--AEDFTFDFNELEKAISPKTKAIILNNPHNPTGQLIPQSD
Q9V0G5          PPYTGLVKFYGGVPVEYRTIEEEGWQPDIDDLRKKITERTKAIAVINPNNPTGALYDKKT
AAT_PYRKO       NAYSLIARQFWREVEVIKTTLDERWIPRVEEIKADL------IIINYPNNPTGRVLSGKE
Q9L0L5          TTYPESIRLAGGVPVEVVADETTGYRVTVEQLEAARTEKTKVVLFVSPSNPTGAVYGEAE
Q979X6          VSYPDIVRLAGGKPVTVSTLE--DYSLDFDLMRKYVTPKTKAIIFNNPTNPTGKVYDEKE
Q8RCV4          VAYHPCTVIAGAKSVFVPTYEEDDFILRADVLEKYITERSKVLILPYPNNPTGAVMPKEA
AAT_SYNY3       LSYPEMVRLAEGTPVIVNTTAATDYKITPEQLRQAITSKSKLFVLNSPSNPTGAVYTPAE
Q8XJT3          VSYPELVKLSDGVPVFIETKKENDFKVTYDELKSVLSENTKAIVINSPNNPTGTVYSKKD
Q97KB8          SNYNSFAKVSGAKIVPFETKIEDNFHLKKEEIIAKITDKTKCIMLSNPCNPTGTVYSYEE
Q8ZDK4          PLWTAAVSLSSGKAIHYMCDEESGWFPDLDDIRSKITPRTRGIVIINPNNPTGAVYSKEL
O59096          VSYAPAVILAGGKPVEVPTYEEDEFRLNVDELKKYVTDKTRALIINSPCNPTGAVLTKKD
```

```
Q98AR6          TTFPAQVLIAGGTPVFVDT-RSNGYVPRPEHIKEAVTERTRAIVVNTPSNPAGAVYDVET
Q8TT07          ACYPNFVKYLGGTPVFVYTSEANGFALEPETVRQCLSPNTKAILINSPSNPGGHVMSPDT
Q9R096          DCYEPMTMMAGGCPVFVTLKASNDWQLDPAELASKFTPRTKILVLNTPNNPLGKVFSRME
Q9RNK6          VSYPDIVRFCGGTPVFIQATIDQDYKITAEQLEKAITQKTRWFIFNSPSNPTGAAYSADE
Q9SIE1          VSYTEQARLADATPVVIPTKISNNFLLDPKDLESKLTEKSRLLILCSPSNPTGSVYPKSL
YD91_METJA      PVTATHTKWYGGEVYNLPLLEENDFLPDLESIPEDIKKRAKILYLNYPNNPTGAQATKKF
Q9CPI6          PLWTAAATLAGGKPVHYLCDEEANWFPDVNDIKSKITKRTKAIVVINPNNPTGAVYSQDL
Q8U097          NAYLLIAKNFEKEVKIIETTLENSWEPEINDNL-----DVDLLILNYPNNPTGKILPREK
Q9CEK7          VSYVDQVKMAGGNPVIVEAKQENNFKVTVEQLEEARTSKSKILLLNSPSNPTGMIYSKEE
AAT_BACSP       VSYPEQVKLAGGVPVYIEATSEQNYKITAEQLKNAITDKTKAVIINSPSNPTGMVYTREE
Q9I015          PLWTAAVSLAGGKPVHYLCDEQANWWPDLEDIKAKITPNTRAMVIINPNNPTGAVYSREV
Q9KQM1          PLWTAAVALSGGKAVHYICDEEADWYPDLDDIRSKITPKTRGIVLINPNNPTGAVYSRDF
Q9A8H2          PIHAFGFIMAGGIIRHVPALSPEEYLSNISRAVKHSVPPPSVLILSYPSNPTAQWVDLDF
O27916          LSYDACVRLSGAVSIPVPLSMDEGFSMSPERVESLITQDTRVIIMNSPSNPTGSVMGKDD
AAT_BACST       VSYPEQVKLAGGVPVYVEGLEQNHFKITPEQLKQAITPRTKAVIINSPSNPTGMIYTAEE
AAB1_RHIME      TSYSDIVHICEGKPVLIACDASSGFRLTAEKLEAAITPRTRWVLLNSPSNPSGAAYSAAD
AATA_RHIME      VSYPEMVALCGGTPVFVPTRQENNFKLKAEDLDRAITPKTKWFVFNSPSNPSGAAYSHEE
Q9HHD3          PPYVGLVKFYGGVPKAYRTVEEEGWQPDIDDMRKKITEKTKAIAVINPNNPTGALYEKKT
Q8Y525          SPYKDQVLNSGGTPIIIPTYEKDDFAINVDILEAAITDKTKALILNSPNNPTGAVFSPET
AAT_RICPR       VSYPDMVALSTGTPVFVNCGIENNFKLSVEALEHSITDKTKWLIINSPSNPTGAGYNCKE
Q8YIC8          VSYPEMVAINGGTPVFVDTKIEDNFKLTAADLEKAITPKTKWLIFNSPSNPTGAAYTQAE
Q9HK41          TEIAEIIKLADGIPIRIPV------ERYVEEMRKYNDDRVRAVFINSPHNPTGYVFGEKQ
AAT_RHILP       VSYPEMVALCGGTRFFVSATQEHNFKLQAADLEKAITPKTKWFIFNSPSNPTGAAYTHDE
Q53951          ------------------------------------------------------------
Q9HKR7          VSYPDIVKLAGGKPVPVRT--TDDYDIDLDEMRKFVTPRTRAIILNNPNNPTGKVLSEKE
O25383          VTYPELVKYSGGVSQFIQTDEKSHFKITPKQLKDALSPKTKMLILTTPSNPTGMLYSKAE
Q9CAP1          DSYEATLSMAGAK-VKGITLRPPDFSIPLEELKAAVTNKTRAILMNTPHNPTGKMFTREE
AAT_THETH       VSYPEMVRFAGGVVVEVETLPEEGFVPDPERVRRAITPRTKALVVNSPNNPTGAVYPKEV
Q9RWP3          VSYPEMVALTGAVPVTVPTTPQQGFQLDPDALAAAITPRTRMVILNSPGNPTGAVFPPET
Q9UZ63          VSYVEDAKLAEAKPVRIPLREENNFMPDPDELLELVTKKTRMIVINYPNNPTGATLDKEV
Q8U1F5          VSYAPAVILAGGKPVEVPTYEDNEFRINVDDLKKHVSEKTRALIINSPNNPTGAVLTKKD
Q08415          DCYEPMTMMAGGCPVFVTLKASNDWQLDPAELASKFTPRTKILVLNTPNNPLGKVFSRME
AAT2_BACSU      PEYLSGITMARAELYEMPLYEENGYLPDFEKIDPAVLEKAKLMFLNYPNNPTGAVADAAF
Q9X8S5          DSYAACIAMAGGTRVPVTLRPEGTFRLDLDELRDAVTDRTRLLLLNTPHNPTGTVLTRDE
Q939K8          PGYEPLITLARAEPIYIDT-TSNGFVLTPEIIEAEHGDQVKAIILNYPSNPTGVTYNREE
Q9RAT0          PGYEPLITLAGGSLVEIDT-RANDFVLTPEMLDQAIIEKVKAVILNYPANPTGVTYNREQ
Q9CJE0          PGYEPLITLAGGSLVEIDT-RANDFVLTPEMLEQAIVEKVKAVILNYPANPTGVTYNRGQ
Q972A2          YSYAEVVKMLGGVPVYVKMKESTGFSLNLSELESKINKKTKMIVLNNPHNPTGMVFDPIE
Q9XBE6          MIYPPMIELLGRRVVRIPTSPADGFRLHLSDLRSRLTQRSRVVLVNSPGNPTGRVSSEDE
Q9HRM6          VSYEAMAKLAGGELNRVDL-APHDFQLELDDLADAVSDDTELLVVNSPSNPTGAVYSRAA
Q92JE7          VSYPDMVALSTGTPVFVNCGIENNFKLSAEALERSITDKTKWLIINSPSNPTGASYNFEE
Q9AA68          DAYLPLVRRAGGVPRLVKL-SPPHWRFERAMLEAAFSNRTRMVVLNSPLNPAGVVAPDED
Q8U821          VSYADIVKFAGGVPVPVPCFEQAGFKLRAEDLEAAITPRTKWLFLNFPSNPTGAACSRKE
Q92D16          PGYEPLIKLNKARPVKVDT-TESNFKLTPEQLRAHITPKTKALIIPYPSNPTGVSLSKQE
O28650          VSYKPLVEICGGEAVTIPT--APEFRLTYEMLMQYRDAKAKILVLNYPANPTGVSYSKKE
AAT_THEMA       VSYIPQIILAGGTVNVVETFMSKNFQPSLEEVEGLLVGKTKAVLINSPNNPTGVVYRREF
Q8YP73          PVYRTSTIFAGGEAFSMPLKAENKFLPDLDLIPEEVARKAKMLWINYPNNPTGALATLEF
O53870          DSYSPVVAMAGAHRVTVPLVDGRGFALDADALRRAVTPRTRALIINSPHNPTGAVLSATE
Q8TR00          VSYNALTEILNGKVVSVPL--AEDLTMKPDAVLERITPKTKALILNSPSNPTGAVSSRAD
Q97I35          VTYPELIKLNDGISVIINTTEENHFKLSVDDLENAYTSKTKAILINSPSNPTGTVYTETE
Q8XGH1          PLWTAAVSLSSGKAVHYLCDESSDWFPDLDDIRAKITPRTRGIVIINPNNPTGAVYSKEL
Q8UDD1          VSYPEMVAICGGTPVFVNATLEDNFKLKPEALEKAITPKTKWFVFNSPSNPSGAAYSHDE
Q54188          ------------------------------------------------------------
Q8Y606          VTYPEQVKLAGGIPVFVETGFDADFKISAADFEKAITKKTKAIVLNSPNNPSGMCYTKEE
```

```
Q8TPT6          VTYDACIRFSGANTVWVPTVPERGFL--PDNFAEYINDKTKLIVVNSPGNPTGGVFGKKT
Q98H83          PIYRSMIEFTGAAPIPVPMREENGFAFSAEETLALITSKTRLLILNSPANPTGGVTPRAE
O31665          PDYWSGVTLAKAKMEMMPLVKDRAFLPDYSSITAEIREQAKLMYLNYPNNPTGAVATSEF
O28151          VSYEACILMAGAKPVWVPH----EEGFEDAPIEDYITSNTKMIVVNTPSNPLGVVYPKEF
Q8Y8A4          PGYEPLITLNKAHPVKVDT-TETNFKLTPEQLKAHITPKTKALIIPYPSNPTGVTLSKDE
Q92AB1          VTYPEQVKLAGGVPVFVETGFDADFKISATDFEHAITEKTKAIVLNSPNNPSGMCYTKDE
O54170          SLYADLVVLAGGTVDFVPL--APDLHWDLDALAA-ALPGAAMMIFSNPSNPTGIVHRREE
Q9A830          VSYPDMTLLAGGTPVSVETTAESGFKITPEALEAAITPKTKWLIINSPSNPSGGAYSRAE
Q8XRN7          PNVVEIPKILGAHVETVPLYGAHGWTLDVDRLLAALTPDTRLLAINSPNNPTGWVMSREA
Q8W360          ETYQTCIELARGVPVYVPL-DPPSWTLNEDKFLKSFTNRTKAVVLNSPHNPTGKVFSREE
Q929C3          SPYKDQVLNSGGTPIIIPTYEKDGFAINVDILEAAITNKTKALILNSPNNPTGAVFSPET
O86587          PLWTAVTTLAGGKAVHYLCDEQAEWYPDLADMEAKITDRTKAVVIINPNNPTGAVYPKEI
O59044          PPYTGLVKFYGGKPVEYRTIEEEGWQPDIDDLRKKISERTKAIAVINPNNPTGALYDKKT
Q943I5          VSYPEMATLAGATPVILPTSISENFLLRPELLASKINEKSRLLILCSPSNPTGSVYPKEL
Q97TA8          PGYEPIVNLVGAEIVEIDT-TENGFVLTPEMLEKAILDKLKAVILNYPANPTGITYSREQ
AAT_STRVG       TTYPESIRLAGGVPVDVVADETTGYRVSVEQLEAARTERTKVVLFVSPSNPTGSVYSEAD
Q8RQG1          ------------------------------------------------------------
Q8XQJ3          VSYPEIVQLCGARSVIVPCGAESGFKLTPGALEAALSARTRWLILNSPSNPTGAVYTAQE
Q8TQ40          PVYKIGTLFAGGEPYSLPLKAENSFLPDLDSIPADILKRAKLFFFNYPNNPTSATADMKF
AAT_RICCN       VSYPDMVALSTGTPVFVNCGIENNFKLSAEALERSITDKTKWLIINSPSNPTGASYNFEE
Q9RWJ7          DVYVPQAELAGARAVPVPLHPEEGWSLDLAAVRAAITPRTQALLVNTPHNPTGLVFSWAE
Q97GI7          PAYASCTKLLEGEVITYGLY-GSEFSIDFNELEKIKNEKPKFMVLSYPSNPTGTVISKED
O95335          IPLR-------SKPVYGKRWSSSDWTLDPQELESKFNPKTKAIILNTPHNPLGKVYNREE
Q98I67          VSYPEMVAICGGTSVFADTSIENGFKLTAEVLEKAITPKTKWLLMNSPSNPSGAAYTQAE
AATC_RHIME      PIHAFGFLMAGGVIRSISVEPDESFFPPLERAVRHSIPKPLALILNYPSNPTAQVATLDF
AAT_STRGR       TTYPESIRLAGGVPVEVVADETTGYRVSVEQLEAARTEKTKVVLFVSPSNPTGAVYSEAD
Q8YA73          PDYLSGVVLGEVQFEKMPLIAENDFLPDFTKIPEDIAEKTELMYLNYPNNPTGAVATADF
Q8ZI88          PSNFLNVELLGGKPISVELKAENNFQIDIKDFENKITEKTKMVILTNPNNPTGTVLRRES
O67864          PIHYYAPIIAGGEVHSIPLNHQEEFLRRLYEIVKTAMPKPKAVVISFPHNPTTITVEKDF
Q93RH7          PDYLSAVALGQVDCETYPLLAENDFLPDLTAIPEESARRAKFIYINYPNNPTGAVATPAF
Q8R7H1          VSYAPCVILTRGVPVFIPTDEKNNFILTPDDLRSKITSKTKALILLYPNNPTGAIMKKED
Q9A0S0          VSYSDQVKMAEGQPIFVQGLEENQFKVTVDQLERARTSKTKVVLINSPSNPTGMIYGAEE
AAT_SULSO       YSYAEVVKLLGGKPINLKWSREEGFSIDVDDLQSKISKRTKMIVFNNPHNPTGTLFSPND
Q8ZW57          PGYFEIANVFDIPLGFVET--GDGWVPQFAERGVYV--------VNYPNNPTGAVLPRHK
Q982E0          TTFPAQVLIAGGKPIFVDT-RSNGYVPRIEDIEAAITERSRAIVINTPNNPTGSVYDTQT
Q982E3          PTFASQILLAGAKPVFVDS-RPSTYIPNIGAVRDALTQRTKAIIVNSPNNPTGIIYDPTT
O66630          PVYRIGAIFAGGTPYTVPLKEENNFLPDLDSIPEDVAKKAKIIWINYPNNPTSAPPTLEF
AAT2_METJA      PCYKNFIRFLGAKPVFCD--------FTVESLEEALSDKTKAIIINSPSNPLGEVIDR--
O29838          PVYYSSTLLADGVPYEMPLKEENKFLPDFQLIPDEIARKAKIMFLNYPNNPTAAVAPKEF
Q9HQK2          ACYPNFVRLADGVVRTVGLAPDAGFQPAVSDYDAAIGDDTAAMLLNSPGNPTGAVIDGES
AAT1_BACSU      VSYPEQVKLAGGKPVYVEGLEENHFKISPEQLKNAITEKTKAIVINSPSNPTGVMYTEEE
Q9YE99          FGYRPIVEYLGGRVEWVRAPPSLGFQPDEERLKEAFTRDVKAVVLVSPDNPTGRLLSTES
Q9R6Q3          VSYVDQVKMAGGNPVIVEAKQENNFKVTVEQLEKARTSKTKILLLNSPSNPTGMIYSKEE
Q97AE8          TEIAEIIKLAEGVPIRLPV------ENYIEEMQKYDDDKVKAVFVNSPHNPTGLVFTPKQ
O30304          LSYFTYAKMCGGQIVQLRTH-NNGFLPDVEKLNEIIDRNVSVIFLNYPNNPTGAVIDEKD
Q92QJ6          PAYRNIMAALGLEVVEIEANAEAGFTLTPDSLEGRAGKPLKGVLLASPANPTGTVTGKAQ
AAT_PYRHO       VSYAEDAKVAEAKPVRIPLREENNFLPDPNELLEKISKNTRMIVINYPNNPTGATLDKEL
Q98KW9          ATYPNTFSAAGASFTVVETPAEDGFQPRADMIRAALRPNTRAILINTPNNPTGAVYSRER
Q8VS38          VSYVDMVSLFGGLPVVIEC--GQGLKLTPELLKNKVTKKTKWLILNSPNNPAGVVYTYDE
Q8XMH8          VAYKGCTAFTGATAKTIDLRACDDFKLTPELLEEAITEKTKVVIIPFPNNPTGAIMNKEE
AAT_PYRAB       VSYAPAVILAGGKPVEVPTYEENEFRLNVDELKKYVTEKTKALIINSPCNPTGSVLKKKD
Q8VS39          VSYVDMVSFFGGLPVIAEC--KDNFKLTSELLRSKITEKTKWLILNSPNNPAGVVYTYDE
Q9KE01          PDYWSGIAIAGGSMYGMPLKKELGFHPDLRGIPAPVLYEAKLMFLNYPNNPTGAVATEEL
O05237          VAYDALVSLAGGIPVHVHTTADKGFKATAADFEAAVTEKTKAILICSPSNPTGSVYSKEE
```

```
AAT_AQUAE       VTYPEQIRFFGGVPVEVPLKKEKGFQLSLEDVKEKVTERTKAIVINSPNNPTGAVYEEEE
Q8REF4          SNYKSFLDIAGAKIIPIPTDIKNDFALKKEEIQKLITSKTKAILYSNPCNPTGKVYTEEE
Q9PPF7          VSYPEMVKFAGGKPVFIEGLEENGFKITAEQLKKAITAKTKVLMLNSPSNPVGSIYSKEE
Q8Y1I0          PCNRHFVPAFDGVARLVPSGPQTRFQLSAEQVEANWTDRTQGVLLASPSNPTGTSILPDE
Q8RR70          VSYPEMVKLAEGTPVILPTTVETQFKVSPEQIRQAITPKTKLLVFNTPSNPTGMVYTPDE
Q9HRX4          VSYDALTRMAGGNPVGLPL--RDDLTLAPETVEDHITDDTAAFVVNSPANPTGAVQSPAD
Q97M25          VGYEPVINFCGAKPVYMDT-SSNNFILNAEILEKYLTPKTKCLILCYPCNPTGSAMDKNA
Q9P9M8          PPYTGLVKFYGGKPVEYRTIEEEDWQPDIDDIRKKITDRTKAIAVINPNNPTGALYDKKT
AAT_THEAQ       VSYPEMVRFAGGVPVEVPTLPEEGFVPDPERVRRAITPRTKALVVNSPNNPTGVVYPEEV
Q8XXV2          DSYLPSIELAGGKAVPI-TLEAPDYRIPFDKLAAAITPRTRLIMLNTPHNPTGTVWHADD
Q8TS80          VAYVPSVILAGGKPVIVSTSRDDDFSLTAEALKPAITSKTKAIILNFPNNPTGAIMEQEG
Q8VXZ8          DSYEATLSMAGAK-VKGITLRPPDFSIPLEELKAAVTNKTRAILMNTPHNPTGKMFTREE
Q8XRB5          VSYPDMVLACDGTPITLPCPEHDGFKLTPARLEAAITARTRWLILNSPSNPTGASYTLDE


1gdeA           LEEIADFVVEHDLIVISDEVYEHFIYDDARHYSIASLDGMFERTITVNGFSKTFAMTGWR
Q8Y0E8          LEAFAEVLRNPRLLILADDIYEHIVFDGLRFASFTAVAPLRHRTLTVNGVSKAYAMTGWR
Q9BVY5          LQVIADLCIKYDTLCISDEVYEWLVYSGNKHLKIATFPGMWERTITIGSAGKTFSVTGWK
Q8RAK7          LQDIAEVVEETGIFVISDEVYEKLIYEGEHVSIASLGEKIKELTIVVNGMSKAYAMTGWR
AAB2_RHIME      YRPLLDVLLKHHVWLLVDDMYEHIVYDAFRFVTPARLEGLKDRTLTVNGVSKAYAMTGWR
Q55453          LRAVNQLCQERGIYHIHDEAYDYFAYDQTPIFSPEAMGDSGGHTISLYSFSKAYGMAGWR
AATB_RHIME      YRPLLDVLLKHHVWLLVDDMYEHIVYDAFRFVTPARLEGLKDRTLTVNGVSKAYAMTGWR
Q8YTF2          FEEIVAFARKYEILLVHDLCYAELAFDGYQPTSLLEIPGAKDIGVEFHTLSKTYNMAGWR
Q9KC79          LAALGEVCLKHNVLIISDEIYEKLVYDGAKHTSVAEISPQLENTVVINGVSKSHSMTGWR
Q92S71          LKAILALAREHGLWIIADEIYALYYYPDGRAPSFLDVMEEDDRILFVNSFSKNWSMTGWR
Q93703          LEEILAFAHQYKLIIIADEIYGDLVYNGATFYPLASLSP-KVPIITCDGIAKRWMVPGWR
Q8YMS6          IKALAQVVVDADIYVVSDEIYEKILYDGAQHISIGSLGKIFNRTLISNGFAKAYSMTGWR
Q9HV76          LAGLSGALKARGGHLVVDEIYHGLTYGVDA----ASVLEVDDDAFVLNSFSKYFGMTGWR
Q8U3E6          AKAIADVAEDYNIYILSDEPYEHFLYDDAKHYPMIKFAP--DNTILANSFSKTFAMTGWR
Q9HUI9          WEALAELCMAHDLWMISDEVYSELLFDGE-HVSPASLPGMADRTATLNSLSKSHAMTGWR
Q8ZVJ5          AKAVVDLAEDYDFWIITDEAYKTLIYEGSHIYLYKLAP---DRTISINTFSKDPAIPGWR
AAT_METEX       ------------------------------------------------------------
Q9ZLG5          LEALGEVLKDTKVWVLSDEIYEKLVYKGEFVSCAAVSEEMKKRTITINGLSKSVAMTGWR
AAT1_METJA      IKGLAEIAEDYNLIIVSDEVYDKIIYDKKHYSPMQFTD----RCILINGFSKTYAMTGWR
Q8XJ54          RDELIEIIKENDILVLTDEIYSSLCFEEE-YYSVAQCKDIKEKIIYVSGFSKMFSMTGLR
Q8UFR3          LKRLASYCESRDMAFISDEIYHGLTFVGEETSALEITDS----AVVINSFSKYYCMTGWR
Q9W6U2          LQMIADLCIKHDVLVISDEVYEWLTYDGAKHVKIASLPGMWERTITIGSGGKTFSATGWK
Q9K7L1          LEAVAKVISGNDLIVFSDEIYAELTYDGT-HVSLASMDGMRERTVLISGFSKAFAMTGWR
O66737          LKELAEYCEEKGMYFISDEIYHGLVYEGREHTALEFSD----RAIVINGFSKYFCMPGFR
Q9Y9P0          VEAIHDIARRRGLIILADEIYDNFLYTEKPFKSTLSLPDWRENLVYVNGFSKTFSMTGWR
Q99V44          VLNIVNVLKKYPIFIISDEIYAENTFSG-KHVSFAEFEDIRDQLILIGGLSKSHSATGIR
Q8RA61          LEEIARVVEEANIFVISDEIYEKLIYEGEHVSIASFGEKIKELTIVVNGISKAYAMTGWR
PATA_BACSU      LKSIAALLKGRNVFVLSDEIYSELTYDRP-HYSIATY--LRDQTIVINGLSKSHSMTGWR
Q98NB8          LSALVNTAEALGIAVISDEIYHRLAYGAPDTTALA----FGNSVTVINSFSKYYCMTGWR
Q97PQ9          LLAIGNWAVENDILILADDIYGRLVYNGHEFTPISSLEAIRKQTVVINGVSKTYAMTGWR
Q9KAU1          LERLNQYLVATEIHVLYDDPYSQLIYDGNVPNPFAAI----ERLFYISSFSKDLGLAGER
Q9V0G5          LEEIINIAGEHDIPVLSDEIYDLMTYEG-KHISPASLTK-DVPVIVMNGLSKVYFATGWR
AAT_PYRKO       IRGLLDVAEENGVKVLSDEVYAELSFT-----RFTPARELYENVVTVKGFSKLYSMTGFR
Q9L0L5          TEAIGRWAVEHGLWVLTDEIYEHLVYGDAVSVSLALLPELRDKCIVVNGVAKTYAMTGWR
Q979X6          IKSLVDFALEYGLYIVSDEIYEDLIYNG-KLISPASYSEMWGKSITLNGFSKGYAMTGWR
Q8RCV4          LEEIAKVVEKHDLIVVTDEIYSELVYGGFKHTSFASLPGMWERTITINGFSKSYAMTGWR
AAT_SYNY3       IRALAAVILEYDLYVVSDEIYERILYDGTEHLSIGAVDEIFQRTIISNGFAKSYSMTGWR
Q8XJT3          LEVIAKFAEENDLIIISDEIYEKLIYGKEEHISIASLEDAFKRTVVINGFSKAYAMTGWR
Q97KB8          MRMLGEIAKEYNLFIISDEVYRQFIFDNVPYTSSIHLTDILDRVILVDSISKHYSACGAR
```

```
Q8ZDK4          LLEIVEIARQNDLIIFADEIYDKILYDDAQHHSIAALAPDL-LTVTFNGLSKTYRVAGFR
O59096          LEEIADFVVEHDLIVISDEVYEHFIYDDARHYSIASLDGMFERTITVNGFSKTFAMTGWR
Q98AR6          LMAIAQLAVSHNLWIIFDECYGDFVHEDHTHHPIVSVAPIRARALIVSSFSKSLALTGWR
Q8TT07          LQGLAAIADEKGIPVVSDEIYQGLIYSGEEHSILEYTKNAF----VLNGFSKLYAMTGWR
Q9R096          LELVANLCQQHDVVCISDEVYQWLVYDGHQHVSIASLPGMWDRTLTIGSAGKSFSATGWK
Q9RNK6          IKSLAEVLRRHHVWILSDDIYEHIVFDNFRFATIAEVAPLFDRTLTANGCSKAYAMTGWR
Q9SIE1          LEEIARIIAKHRLLVLSDEIYEHIIYAPATHTSFASLPDMYERTLTVNGFSKAFAMTGWR
YD91_METJA      YKEVVDFAFENEVIVVQDAAYGALVYDG-KPLSFLSVKDAKEVGVEIHSFSKAFNMTGWR
Q9CPI6          LLEIIEVARQHKLIIFADEIYDKILYDDAVHHHIAALAPDI-LTVTFNGLSKAYRVAGFR
Q8U097          LKELVEVAEEKGIKILSDEIYAEISFK-----SFTPVRELYENTVTVKGFSKLYSMTGFR
Q9CEK7          LTAIGEWALAHDLLILADDIYHRLVYNGAEFTAISSLDEIRKRTTVINGVSKTFAMTGWR
AAT_BACSP       LEDIAKIALENNILIVSDEIYEKLLYNGAEHFSIAQIEEVKAQTIVINGVSKSHSMTGWR
Q9I015          LEGMVELARQHNLVLFSDEIYDKILYDGAVHVSTASLAP-DVLCLTFNGLSKSYRVAGFR
Q9KQM1          LLEIIEIARKHKLMIFADEIYDKVLYDGAVHTSIATLAD-DVLVVTFNGLSKAYRVCGFR
Q9A8H2          YKDAVALAKKHDLLVISDVAYGEIYFDNNPPPSILQVDGAKDIAVEVNSLSKTYAMAGWR
O27916          VKGIAEIAEDNDLIIISDEIYEKIIYDG-KHYSPAQFT---DNALIVNGFSKTYAMTGLR
AAT_BACST       LKALGEVCLAHGVLIVSDEIYEKLTYGGAKHVSIAELSPLKAQTVIINGVSKSHSMTGWR
AAB1_RHIME      YRPLLEVLLRHHVWLLVDDMYEHIVYDGFRFVTPAQLEGLKNRTLTVNGVSKAYAMTGWR
AATA_RHIME      LKALTDVLMKHHVWVLTDDMYEHLTYGDFRFATPVEVEGLYERTLTMNGVSKAYAMTGWR
Q9HHD3          LQEIIDLAGEYDLPIISDEIYDLMTYEG-KHVSPGSLTK-DVPVIVMNGLSKVYFATGWR
Q8Y525          FEKIANLAKKYDFFILSDEVYDGFSFYED---FVPMAKFAPDHTITFGSMSKNFAMTGWR
AAT_RICPR       LENIAKTLRKYNVNIMSDDIYEHITFDDFKFYTLAQIAPDLERIFTVNGVSKAYSMTGWR
Q8YIC8          LKSLTDVLVRHHVWILTDDMYEHLVYGDFVFTTPAQVEPLYDRTLTMNGVSKAYAMTGWR
Q9HK41          ISDIIDFAESKGIFIVSDEAYEDVIFDGLKHLSPGSL---YDDTISLFSMSKSYAMSGLR
AAT_RHILP       LKALTDVLMNPQVWVLTDDMYEHLTYGDFKFVTPVEVEPLYDRTLTMNGVSKAYAMTGWR
Q53951          ------------------------------------------------------------
Q9HKR7          IRDLVDFALENDIYIVSDEIYEDLIYEGSLYSPASMGKEAFEHTITLNGFSKGYAMTGWR
O25383          LEVLGEVLKDTKVWVLSDEIYEKLVYKGEFVSCAAVSEEMKKRTITISGLSKSVAMTGWR
Q9CAP1          LETIASLCIENDVLVFSDEVYDKLAFEMD-HISIASLPGMYERTVTMNSLGKTFSLTGWK
AAT_THETH       LEALARLAVEHDFYLVSDEIYEHLLYEGE-HFSPGRVAP--EHTLTVNGAAKAFAMTGWR
Q9RWP3          LRAVADLATQHGLMIVTDEIYEHLVYDAEQ-VSIGTYAP--EHTLTINGASKAYAMTGWR
Q9UZ63          AKAIADIAQDYNIYILSDEPYEHFIYDDAKHYPMLKFAP--ENTILANSFSKTFAMTGWR
Q8U1F5          LEEIADFANEHDLMIISDEVYEHFIYDGAKHYSIAALDGMFGRTITVNGFSKTFAMTGWR
Q08415          LELVANLCQQHDVVCISDEVYQWLVYDGHQHVSIASLPGMWDRTLTIGSAGKSFSATGWK
AAT2_BACSU      YAKAAAFAKEHNIHLIHDFAYGAFEFDQ-KPASFLEAEDAKTVGAELYSFSKTFNMAGWR
Q9X8S5          LAAIAELAVERDLLVVTDEVYEHLVFGTAEHIPLASFPGMRERTVTIGSAGKTFSFTGWK
Q939K8          VKAIADAVKKYSIFVISDEIYSELTYGET-HVSIAEFAR--DQTILINGLSKSHAMTGWR
Q9RAT0          IKDLAEVLKKHEVFVIADEVYSELNYTDQPHVSIAEYAP--EQTIVLNGLSKSHAMTGWR
Q9CJE0          IKALAEVLKKHEVFVIADEVYSELNYTDQPHVSIAEYAP--EQTIVLNGLSKSHAMTGWR
Q972A2          IEKLMEITKEKKVLLLSDEIYDYFIYEG-KMKSVLEDPDWRDYVIYVNGFSKTFSMTGWR
Q9XBE6          LADLCAFAVEHGLYVVHDEVLDRFAYGIEHRSVVALDHQ--GVGIAVNGLSKRFGMSGWR
Q9HRM6          MEGVRDLAVDHDITVISDEIYQRVNYGPA-HVSLAGLDGMFERTVTINGFSKAYSMTGWR
Q92JE7          LENIAKVLRKYHVNVMSDDIYEHITFDDFKFYTLAQIAPDLKRIFTVNGVSKAYSMTGWR
Q9AA68          LALLAEVCVRHDVVAVCDEVWEAVVFDGRRHRPLMSFPGMRERTVKIGSAGKLFGMTGWK
Q8U821          MAAIAEVMLRHHVWILTDDIYEHLVYDGFEFGTIADVEPLYDRVLTMNGVSKAYAMTGWR
Q92D16          LRDLAEVLKETGIFVIADEIYSELTYHEE-HVSIAPM--LRDQTIVINGLSKSHAMIGWR
O28650          LEEIADAVNELDLIVLSDEIYAELTYTG-RHVSMAALNGMEDRVVIFNGFSKAFAMTGMR
AAT_THEMA       LEGLVRLAKKRNFYIISDEVYDSLVYTDE-FTSILDVSEGFDRIVYINGFSKSHSMTGWR
Q8YP73          FEELVALCQQYSILLCHDHAYSEMAYDGYKPPSVLQIPGAKDIAIEFHSLSKSYNMTGWR
O53870          LAAIAEIAVAANLVVITDEVYEHLVFDHARHLPLAGFDGMAERTITISSAAKMFNCTGWK
Q8TR00          IKALAEIADDHNITIISDEVYEYFIYEGEH----VSPASYSDNVVTINATSKSYSMTGWR
Q97I35          LKAIAEFAKEKDLFIISDEIYEKLIYDGERHVSIASLQDAFNRTVVINGMSKSYAMTGWR
Q8XGH1          LMEIVNIAREHNLIIFADEIYDKILYDDAEHHSIAALAPDL-LTITFNGLSKTYRVAGFR
Q8UDD1          LKALTDVLVKHHVWVLTDDMYEHLTYGDFKFVTPVEVEPLYDRTLTMNGVSKAYAMTGWR
```

```
Q54188          ------------------------------------------------------------
Q8Y606          LIAIGEVAEKHQIYILSDEIYEKLYYGNKADVSIASLDRLYDLTIVINGVSKAYSMTGWR
Q8TPT6          LQCIADLAIDHDLLVVSDEIYEKIIYDRE-HISIGSFDGMQDRTITVNGFSKAYAMTGWR
Q98H83          IEKLVKGLEKHDVAILSDEIYDAMTYDGETHCSLLGYPEIRDRLIVLNGWSKTWAMTGWR
O31665          FEDTVRFAAENGICVVHDFAYGAVGFDGCKPLSFLQTEGAKDIGIEIYTLSKTYNMAGWR
O28151          LKKVRDLAVDKDILVMSDEIYEKIIFEGE-HYSLAAMDGMLERTITINGFSKTYSMTGWR
Q8Y8A4          LSALAEVLKETGIFVIADEIYSELTYHEE-HVSIAPM--LRDQTIVINGLSKSHAMIGWR
Q92AB1          LAAIGAVAEKHQIYILSDEIYEKLYYGNKADVSIASLDRLYDLTIVINGVSKAYSMTGWR
O54170          LEALGKLLDGTDVLVVSDEAYHRLAYPGHEPVSALEIESLRGRTVYVQTFSKTYAMTGWR
Q9A830          LQAIADVLLRHQVWVLTDDMYEHLVFDDFEFTTIAQVEPLYDRTLTMNGVSKGYSMTGWR
Q8XRN7          QQAVLAHCRRHGIWILADEVYERLYYGDRPAPSFLDIAGRDERVICVNSFSKSWLMTGWR
Q8W360          LLIIAQACQKMDCFAITDEVYEYITYDENKHISLASLPGMQERTIITSSLSKTYSVTGEQ
Q929C3          FEKIANLAKKYDFFILSDEVYDGFSFYED---FVPMAKFAPDHTITFGSMSKNFAMTGWR
O86587          VEGILDLARRHGLMVLADEIYDQILYDDAVHHSAASLAPDL-VVLTFCGLSKTYRVAGFR
O59044          IEEIINVAGEHDLVVLSDEIYDLMTYEG-KHISPGSLTK-DVPVIVMNGLSKVYFATGWR
Q943I5          LEEIADIVKKYRLLVLSDEIYEHIIYQPAKHTSFASLPGMWDRTLTVNGFSKAFAMTGWR
Q97TA8          LEALAAVLRKYEIFVVCDEVYSELTYTGEAHVSLGTM--LRDQAIIINGLSKSHAMTGWR
AAT_STRVG       AKAIGEWAAEHGLWVLTDEIYEHLVYGEAKFTSLPVLPALRDKCIIVNGVAKTYAMTGWR
Q8RQG1          ------------------------------------------------------------
Q8XQJ3          LRALAEVLLAHDVLVLSDDIYEHLIFDGARFHTLAQVEPLQSRVLTMNGVSKAYAMTGWR
Q8TQ40          FEKVVEFCKKNDIIAVHDNAYSQMVYDGYDAPSFLAAEGAMDIGIELYSHSKTYNMTGWR
AAT_RICCN       LENIAKVLRKYHVNVMSDDIYEHITFDDFKFYTLAQIAPDLKRIFTVNGVSKAYSMTGWR
Q9RWJ7          LTELVALAREHDLWLISDEVYDE-LYASERPTSLRELAP--ERTFTVGSAGKRLEVTGWR
Q97GI7          NEKLHKIIKDNDIIAVTDEMYSALCYEDD-YYSVSQYEDIREKVIVVSGFSKTFSMTGLR
O95335          LQVIADLCIKYDTLCISDEVYEWLVYSGNKHLKIATFPGMWERTITIGSAGKTFSVTGWK
Q98I67          LRALADVLLKHHVWTLTDDMYEHLTYGDFAFKTIAEVEPLYERTLTMNGVSKAYAMTGWR
AATC_RHIME      YKDVIAFAKKHDIIVLSDLAYSEIYFDDAPPPSVLEVPGATDVTVEFTSMSKTFSMPGWR
AAT_STRGR       AEAIGRWAVEHGLWVMTDEI----------------------------------------
Q8YA73          FEETVAFAKNHNIVVAHDFAYGGIGFDGKKPISFLETNGAKEVGIELYTLSKTYNMAGWR
Q8ZI88          LQAIADFIIAHDLILVVDQAFEDAIFDEIEFISIASLPGMWERTVSVFSFSKGMGLSGFR
O67864          FKEIVKFAKEHGLWIIHDFAYADIAFDGYKPPSILEIEGAKDVAVELYSMSKGFSMAGWR
Q93RH7          YESLVDWAKKYEVGVVSDFAYGALGYQGYKNPSFLSTPGAKDVGIELYTFSKTFNMAGWR
Q8R7H1          LEEIVDVIIEKDLIVISDEIYSELTYEG-KHVSIASLPGMKERTILINGFSKAFAMTGWR
Q9A0S0          LRAIGEWAVHNDILILADDIYGSLVYNGNQFVPISTLEAIRRQTITVNGVAKSYAMTGWR
AAT_SULSO       VKKIVDISRDNKIILLSDEIYDNFVYEGKMRSTLEDSDW-RDFLIYVNGFSKTFSMTGWR
Q8ZW57          VKELVDVAE----FIISDEIYRDISFVE-----FTSPLELSPNVAVVYSFSKVFSVPGLR
Q982E0          LTGIAQLAIARDLWIIFDECYADFVHGDEAHRPIVFLPEVRSRTLIVNAFSKSLALTGWR
Q982E3          LRAIGDLAIDHHLWIVSDECYSCFVFAGRHESIVTAHPGVRSRTILVNTFSKELAITGWR
O66630          YKKLVDWAKEYNVIIASDNAYSEIYTGQEKPPSILQVPGAKDVAIEFHSLSKTYNMTGWR
AAT2_METJA      --EIYEFAYENIPYIISDEIYNGLVYEG-KCYSAIEFDENLEKTILINGFSKLYAMTGWR
O29838          IKEAIDFCIDNKIILAHDAAYSEITFDGYKAPSFLEFEDAFEVCVEFNSLSKTYNMTGWR
Q9HQK2          LSALVALADRTDTAVVSDEIYHGLAFDAAAHSVLEYTDDAF----VIDGVSKRYGMTGWR
AAT1_BACSU      LSALGEVCLEHDILIVSDEIYEKLTYGGKKHVSIAQLDRLKEQTVIINGVSKSHSMTGWR
Q9YE99          AKLVADLAVDTGAWIVYDEAYKTLVFEGEHVYLYKLAP---DNTISINTFSKDPGFPGWR
Q9R6Q3          LTAIGEWAVAHDLLILADDIYHRLVYNGAEFTAISSLDEIRNRTTVINGVSKTFAMTGWR
Q97AE8          IDGIISFAESKGIFIVSDEAYEDVIFDGREHVSPGSK---YDNTISLFSMSKTYAMSGLR
O30304          ARAVVEIAADNKAIVVSDEIYDQIYYDVKP----TSLAG-YENVVCVNGFSKSLSMTGWR
Q92QJ6          LKALADYCRDHSIAFISDEIYHGLTFAGEETTALEVADD----AVVINSFSKYYCMTGWR
AAT_PYRHO       AKTIADIAEDYNIYILSDEPYEHFIYEDAKHYPMIKFAP--ENTILANSFSKTFAMTGWR
Q98KW9          LEQLAQICREHDLWLLSDEVY--WTLGGGEHVSPRSLPGMAERTLVINSMSKSHGMTGWR
Q8VS38          LKAIAQVLLEYHVNIITDDIYEH-IYDEKFFTIAQVEPKLYNRVFVVNGVSKAYAMTGWR
Q8XMH8          LQKIVDVLKDKDVIIISDEIYAELSYDED-HVSIASFQEVKEKTIVINGFSKAYAMTGWR
AAT_PYRAB       LEEIADFAVEHDLIVISDEVYEHFIYDDVKHYSIASLDGMFERTITVNGFSKTFAMTGWR
Q8VS39          LKKIAQILLEYHVNIITDDIYEHIVYDGKFFTIAQVEPRLYDRVFVVNGVSKAYAMAGWR
```

```
Q9KE01          FAEAIELAEEYDICVVHDFAYSAIGFDGQKPLSFLQVEGAKNVGIEMITLSKNYNMAGWR
O05237          LNEIAEFAKKHDVIVLADEIYAELTYDEE-FTSIAALPGMKERTVVISGFSKAFAMTGWR
AAT_AQUAE       LKKIAEFCVERGIFIISDECYEYFVYGDAKFVSPASFDEVKNITFTVNAFSKSYSMTGWR
Q8REF4          VKLLANLAAENDLFVIADEPYREFIYDDNKHYSLLDIEKAKENVIIIDSVSKHYSACGAR
Q9PPF7          LTQIAKVLEGTQITVLSDEMYEKLRYDGFDFVAFASVSKDLKRTVTINGLSKCGAMPGWR
Q8Y1I0          LRRIVETVRGRGGFSIVDEIYQGLSYDQAPVSALSFGDD----VVTINSFSKYFNMTGWR
Q8RR70          VRAIAQVAVEAGLWVLSDEIYEKILYDDAQHLSIAASPEAYERSVVCSGFAKTYAMTGWR
Q9HRX4          MRAFARIADEHDVLCISDEVYEHIVFEGEHRSPMEFADT--DNVVVVNACSKTYSMTGWR
Q97M25          LTEIVNLLKNKDIFVISDEIYSELTFKK-RHFSIAGFEDMRNKTILLNGVSKSHSMTGWR
Q9P9M8          LEEILNIAGEYEIPVISDEIYDLMTYEGE-HISPGSLTK-DVPVIVMNGLSKVYFATGWR
AAT_THEAQ       LRALAEMALQHDFYLVSDEIYEHLIYEGA-HFSPGTLAP--EHTITVNGAAKAFAMTGWR
Q8XXV2          MRRLAEIVAPTDVLLLSDEVYEHMVYDGVPHASVSRIPELARRAFVVSSFGKTYHVTGWK
Q8TS80          MEDIADLVVENDLFVISDEVYECLTYGGT-HVPFSSLEGLKDRTVMLNGFSKAYAMTGLR
Q8VXZ8          LETIASLCIENDVLVFSDEVYDKLAFEMD-HISIASLPGMYERTVTMNSLGKTFSLTGWK
Q8XRB5          YRALADVLARHHVLVMTDDIYEHIRFDTVRTHLLNAAPELRDRTLVINGVSKTYAMTGWR


1gdeA           LGFVAAPSWIIERMVKFQMYNATCPVTFIQYAAAKALKDERSWKAVEEMRKEYDRRRKLV
Q8Y0E8          VGYGCGPKPLIDAMANVQMQVNSHTASISQAAAIAALEGPQD--ELARRRAIFEQRRDSL
Q9BVY5          LGWSIGPNHLIKHLQTVQQNTIYTCATPLQEALAQAFWIDIKRMDDPELPKELEVKRDRM
Q8RAK7          IGYTASSLDVAKVMANIQSHTTSNPNSIAQYASVTALTGDGV--AIKRMVEEFNKRRLYA
AAB2_RHIME      IGYAGGPRALIKAMAVVQSQATSCPSSVSQAASVAALNGPQDF--LKERTESFQRRRNLV
Q55453          VGYMVIPLELLLAVKKIQDTNLICPVVVSQYAALACLRVGKNYSAQFLPEMAACRQQLLE
AATB_RHIME      IGYAGGPRALIKAMAVVQSQATSCPSSVSQAASVAALNGPQDF--LKERTESFQRRRNLV
Q8YTF2          VGFVVGNRHVIQGLRTLKTNLDYGIFAALQTAAETALQLPDIY--LHEVQQRYRTRRDFL
Q9KC79          IGYAAGPEALIQAMTNLASHSTSNPTTSAQYGAIAAYTEDDG--SVEKMRVAFEERLHTI
Q92S71          VGWIVAPPAMGQVLENLIQYSTSGVAQFMQRGAVAALDEGDGF-VEENIAKAKRNRDTLC
Q93703          LGWLIIHNHVKNGIVALSQ-KIVGPCSLVQGALPKILRETPEDYFVYTRNVIETNANIVD
Q8YMS6          LGYLAGPVDIIKAASSIQGHSTSNVCTFAQYGAIAALEDSQDC--VEEMRQAFAKRRQVM
Q9HV76          LGWLVAPTAAVPELEKLAQNLYISAPSMAQQAALACFEPT-TIAILEERREEFARRRDFL
Q8U3E6          LGFVIAPTQIIREMIKLHAYIIGNVASFVQVAGIEALRSKESWKAVEEMRKEYNERRKLV
Q9HUI9          VGWVVGPAALCAHLENLALCMLYGSPEFIQDAACTALEAPLP--ELEAMREAYRRRRDLV
Q8ZVJ5          LGYVYGPPEVMPKIKLVNEEMVYCPPSFAQRLVAIYLRSEARMRYIREVVEIYRQKRDVA
AAT_METEX       ------------------------------------------------------------
Q9ZLG5          MGYAASKDKLVKLMSNLQSQCTSNINSITQMASIVALEGLVD-KEIETMRQAFEKRCHLA
AAT1_METJA      IGYLAVSDDLINNMIKIHQYSFACATTFAQYGALAALRG--SQKCVEDMVREFKMRRDLI
Q8XJ54          IGYVACPKKIYDQIIKVHQYNSSCATSISQWGALEGLKSCMN--DVENMKESFKERMNFT
Q8UFR3          IGWMVLPENLVRPVECLAQSLYISPPELSQLAATAAFSAAEE---LDVYRESYRTNRDFL
Q9W6U2          VGWAISSGHIIKHMKTIHQNTVYHCATPAQEAVARGFEREYEVSYFQQLPAMLHHKKNKL
Q9K7L1          LGYVCAPDDILSAMLKIHQYSLMCAPTMAQHGALEALETGMD--DVHRMVQSYRQRRNFV
O66737          IGWMIVPEELVRKAEIVIQNVFISAPTLSQYAALEAFDYEY----LEKVRKTFEERRNFL
Q9Y9P0          LGYVVLRREVIPKALDLAVTIYSCPPSIAQKAGVAALRGD--WGPVREMVEEFRSRARIL
Q99V44          IGFLLGPQYLIDKLTFMHAYNCICANVPAQIACITALNEGLE--APKYMNEAYVERRNYL
Q8RA61          IGYTASSLEVAKVMSNIQSHTTSNPNSIAQYASVAALQGGEE--EIEKMKEEFNRRRLYM
PATA_BACSU      IGFLFAPKDIAKHILKVHQYNVSCASSISQKAALEAVTNGLDDALI--MREQYKKRLDYV
Q98NB8          IGWMVLPEELVRPVERIAQSLYISPPELSQIAAIEAFAATEELEAV---KGRYAWNRELL
Q97PQ9          IGYAVGEADIIAAMSKIAGQTTSNPSAVAQYAAVEALSGEQD--TVESMRQAFEERLNTI
Q9KAU1          LGYLAIRPDFPSVEQYMAAFVFANRTLGFGNATVTVQRMISKMDTLTQEQHEYKKRRDAM
Q9V0G5          LGYMSEVREAIDKLARIRL----CPNTPAQFAAIAGLRGPMDY--LEEYMKKLKERRDYI
AAT_PYRKO       LGYAIGERNEIRRIQRFIESTVTCVPPFVQRAGVKALELRDEL--IKEVRRAYLERVRMA
Q9L0L5          VGWVIGPKDVVKAATNLQSHATSNVANVSQAAALAAVSG--NLDAVAKMREAFDRRRQTI
Q979X6          IGYMAAPREIVEAANVIQQQTITCVSSISQYAALRALDDTESPK---KMKDEFKRRRDLI
Q8RCV4          LGYIAAPEHFTKHIAKLHQYAVTAAATMCQYAGIEAMRNGDE--DIIKMREEYDKRRKFL
AAT_SYNY3       VGYLAGELPLIQACSTIQGHSTSNVCTFAQYGAIAALENPQTC--VETMVKAFTERRQVI
```

```
Q8XJT3          IGYAACYNELIKVMNNVQSHMTSNTNSIAQFAALEALNGDQE--TIKNMVKEFSLRRELM
Q97KB8          IGLVASKNKDLMHQILKLCQARLCVSTIEQYAAANLINTMGSY--ITDVKMAYKKRRDIM
Q8ZDK4          QGWMVHAKGYIEGL-EMLASMRLCANVPMQHAIQTALGGYQSISEFIQPGGRLYEQRDRA
O59096          LGFVAAPSWIIERMVKFQMYNATCPVTFIQYAAAKALKDERSWKAVEEMRKEYDRRRKLV
Q98AR6          IGYLAGPKEVINAVNALQSHTTSNPNVIAQHAVLAHLQRGGSDYEGKLRSRLTSARRIGL
Q8TT07          LGYIICPPGCVRAIQKIHQNFFICANSFVQEAGIAALKGSQEH--VVEMVQIYNMRRQYM
Q9R096          VGWVMGPDNIMKHLRTVHQNSIFHCPTQAQAAVAQCFEREQQHSYFLQLPQAMELNRDHM
Q9RNK6          IGFAGGPTPLIKAMAKLQSQSTSNPCSISQAAAVAALNGPQDF--LQGWSDDFARRRNLV
Q9SIE1          LGYLAGPKHIVAACSKLQGQVSSGASSIAQKAGVAALGLGKAGETVAEMVKAYRERRDFL
YD91_METJA      LAFLVGNELIIKAFATVKDNFDSGQFIPIQKAGIYCLQHPEI---TERVRQKYERRLRKM
Q9CPI6          QGWMIAAAGYIEGL-DMLASMRLCANVPMQHAIQTALGGYQSINEFILPGGRLLEQRNKA
Q8U097          LGYAIADKEEIRKIKTFIESTVTCVPPFVQRAGIKALELRDEL--MKKVSREYKRRAELA
Q9CEK7          IGLAVGDPEIISAMTKIASQTTSNPTAVAQYAAIEAFEERDN--SFEIMHAAFEERLNII
AAT_BACSP       IGYAAGNADIINAMTDLASHSTSNPTTASQYAAIEAYNGPQD--SVEEMRKAFESRLETI
Q9I015          SGWVAAQSYIEGLDILANMR--LCANVPAQHAIQTALGGYQSINDLVLPPGRLLEQRNRA
Q9KQM1          GGWMQAQGYIAGLDMLASMR--LCANVPMQHAIQTALGGYQSINELILPGGRLLEQRDRA
Q9A8H2          VGMVVGNARICAALARVKSYLDYGAYTPVQVAAATALNGPQDC--VDEIRGIYKSRRDTL
O27916          IGYVAGCEDIIEELLKVHQYNTACAPSISQYAALAAIRGPQNC--VKDMVDEFRRRRDLM
AAT_BACST       IGYAAGPKDIIKAMTDLASHSTSNPTSIAQYAAIAAYSGPQE--PVEQMRQAFEQRLNII
AAB1_RHIME      IGYAGGPRELIKAMAVVQSQATSCPSSISQAASVAALNGPQDF--LKERTESFQRRRDLV
AATA_RHIME      IGYAAGPLHLIKAMDMIQGQQTSGAASIAQWAAVEALNGPQDF--IGRNKEIFQGRRDLV
Q9HHD3          LGYMAEVREAIGKLARIRL----CPNTPAQKAAIAGLRGPMDY--LEEYMAKLKERRDYI
Q8Y525          LGYMIAPTYLNEAAKIINEGITYSAPTPSQRAAIYALNHSETL--IPLVAETFQKRLEYI
AAT_RICPR       IGYGAGSKALIKAMTIIQSQSTSNPCSISQMAAIEALNGTQDY--IKSNALNFQKKRDLA
Q8YIC8          IGYAAGPIELIKAMDMIQGQQTSGACSIAQWAAVEALNGTQDF--IPANKKIFQARRDLV
Q9HK41          IGYAHTSNEILDRMKKLLRCTINGVNSATQYGAVAALTGPQDY--IGEMRKEYQKRRDII
AAT_RHILP       IGYAAGPIQLIKAMDMIQGQQTSGATSIAQWAAVEALNGTQDF--IPENKKIFEGRRDLV
Q53951          ------------------------------------------------------------
Q9HKR7          IGYFVATEEIVEAANIIQQQTITCASSISQYAALRALDDNES---PARMRSEFRKRRDLA
O25383          MGYAASKDKLVKLMNNLQSQCTSNINSITQMASIVALEGLVD-KEIETMRQAFERRCDLA
Q9CAP1          IGWAIAPPHLTWGVRQAHSYLTFATSTPAQWAAVAALKAPESY--FKELKRDYNVKKETL
AAT_THETH       IGYACGPKEVIKAMASVSSQSTTSPDTIAQWATLEALTNQEASRAVEMAREAYRRRRDLL
Q9RWP3          IGYAGGPREVIAAMNALQSQSTSNASSVSQYAALAALEQEETMRFIDRARTAYRERRDRI
Q9UZ63          LGFVVAPSEIIKEMIKLHAYIIGNVASFIQVAGVEALRSEESWKAVKEMRKEYNERRKLV
Q8U1F5          LGFVVAPSWVIEKMVKFQMYNATCPVTFIQYAAAKALRDERSWKAVEEMRKEYERRRNLV
Q08415          VGWVMGPDNIMKHLRTVHQNSIFHCPTQAQAAVAQCFEREQQHSYFLQLPQAMELNRDHM
AAT2_BACSU      MAFAVGNEKIIQAVNEFQDHVFVGMFGGLQQAASAALSGDPEH--TESLKRIYKERIDFF
Q9X8S5          VGWVTAAPALLTAVRSAKQYLTYVASGPFQYAVAEALALPESY--FAAYRQDMEAKRDLL
Q939K8          IGFILAPQELIGQIVKVHQYLVTSATTMAQKAAIAALTAGAD--DALPMKIEYMKRRDFL
Q9RAT0          IGLIFAARELVAQIIKTHQYLVTSASTQSQFAAIEALKNGAD--DALPMKKEYLKRRDYI
Q9CJE0          IGLIFAARELVAQIIKTHQYLVTSASTQSQFAAIEALKNGAY--DALPMKKEYLKRRDYI
Q972A2          LGYVVAKEKVIKKMAEIAANIYTCPTSFAQKGALAAFE---SFDEVKEMISLFKKRRDIM
Q9XBE6          IGWLASSSAVIAEAAKAHTFFMLAVSHAVQLAAAAALSDPKADGEVSVYAAEIRRRGEVF
Q9HRM6          LGYLAGPEALVDQAGKVQSHSVSSAANFIQRAGVEAIRHTDD--AIDEMVAAFESRRDLL
Q92JE7          IGYGAGSKTLIKAMTIIQSQSTSNPCSISQMAAIEALNGPQDY--IKPNALNCQKKRDLA
Q9AA68          VGFLCAAPPLARALAAAHQFLTFTTPPNLQAGVAWGLDNHRAW--FDDMPANLQRSRDRL
Q8U821          LGYCASGSELITAISNVNGQNGGGITTVTQAAAIAALDGPQDL--LKERAAIYRERRDFV
Q92D16          IGFLLAPEKITTEMLKIHQYSVTCASSISQKAALEAITNGKD--DAFQMRTEYKTRANFT
O28650          VGYVIAPPDIFAGMLKIHQYCMLCAPITGQIGAIEAL---RSLDEVERMRAEYMRRRNFV
AAT_THEMA       VGYLISSEKVATAVSKIQSHTTSCINTVAQYAALKALEVDNSY-----MVQTFKERKNFV
Q8YP73          IGFAVGNAYAIKGLSQVKTNVDSGVFKAIQKAAIAAYATDEV--ELQAVMSVYQSRRDII
O53870          IGWACGPAELIAGVRAAKQYLSYVGGAPFQPAVALALDTEDAW--VAALRNSLRARRDRL
Q8TR00          LGYLAARKEYIAQMNKVHQYIQACANSIAQKAAYAAVTGPKD--SVNAMREEFRKRRDLL
Q97I35          LGYAASGSSFIKLMSHIQAHTTSNANSITQYASVEALNGRQE--ELHSMVTEFEKRRTYM
```

```
Q8XGH1          QGWMVHAKGYIEGL-EMLASMRLCANVPAQHAIQTALGGYQSISEFILPGGRLYEQRNRA
Q8UDD1          IGYAAGPLPLIKAMDMIQGQQTSGASSIAQWAAVEALNGTQDF--IPENKKIFEGRRDLV
Q54188          ------------------------------------------------------------
Q8Y606          IGYAAANKEIIAGMSKLADHLTSNPTANAQYAALEAYVGSQE--VPEKMYQAFEERMERF
Q8TPT6          LGYLTAPPEIFKLLQKIQSHSVSSATTFVQYGGLEALQGPQD--GVKAMVDRFKMRRDIL
Q98H83          MGWSIWPNHLYDKVRKLAVNCWSCVNAPSQFAGIAAIDGPQD--DVDTMMRAFDRRRKVV
O31665          VGFAVGNASVIEAINLYQDHMFVSLFRATQEAAAEALLADQTC--VAEQNARYESRRNAW
O28151          LGYAAAPEWIIKLMNRMQSHSVSHPTSFVQYAGVAALKGDQSF--IKEIVEEFRARRDMI
Q8Y8A4          IGFLLAPEILTQEMLKIHQYSVTCASSISQKAALEAITNGKD--DAFQMRTEYKTRANFT
Q92AB1          IGYAAANKEIIAGMSKLADHLTSNPTANAQYAALEAYVGSQE--VPEKMYKAFEERMERF
O54170          VGYLTGPREVLDAAAQVHRTWNGSLNTAVQHAALAALDLPDG--VVGAMADRYRQRRDLV
Q9A830          IGYAAGPEPLIKAMGKMISQTTSNPCSISQWAALEALNGTQDF--IKPNAKLFQERRDLV
Q8XRN7          LGWMVLPAAVTDDLGKLIEYNTSCAPSFVQEAGVVAVRDGEDFIRGETAR-LRAARDHLV
Q8W360          ------------------------------------------------------------
Q929C3          LGYMIAPTYLNEAAKIINEGITYSAPSPSQRAAIYALNHSETL--IPLVTETFQKRLEYI
O86587          SGWLVVTGRDYLEGLTMLASMRLCANAPAQYAIQAALGGRQSIRELTAPGGRLHEQRDVA
O59044          LGYMSEVREAIDKLARIRI----CPNTPGQFAAIAGLTGSMDY--LKEYMKKLKERRDFI
Q943I5          LGYLAAPKHFVAACGKIQSQFTSGASSISQKAGLAALNLGYAGEAVSTMVKAFQERRDYL
Q97TA8          LGLIFAPATFTAQLIKSHQYLVTAANTMAQHAAVEALTAGKN--DAEPMKKEYIQRRDYI
AAT_STRVG       VGWVIAPQDVIKAATNLQSHATSNVSNVAQVAALAAVSG--NLDAVAEMRKAFDRRRQTM
Q8RQG1          ------------------------------------------------------------
Q8XQJ3          IGFGAGPRWLLEAMEKLQGQQTSGACSISQHAAVAALRGPQDF--IGASRAAFERRRDLV
Q8TQ40          LGFAVGSKALIKGLGKVKSNVDSGVFDAIQIAGIAALSSSQAC--VDDTNKIYEERRNVL
AAT_RICCN       IGYGAGSKTLIKAMTIIQSQSTSNPCSISQMAAIEALNGPQDY--IKPNALNCQKKRDLA
Q9RWJ7          VGWILCPPSLAGGLANVRQVTSFCSPAPLQAAVAEALPLARSQGYYAALRADYAARRALL
Q97GI7          IGYVCAESSFMSSILKVHQYTTTCAPSISQYGALEGLKNCDE--DVQYMKNEFKKRRDYV
O95335          LGWSIGPNHLIKHLQTVQQNTIYTCATPLQEALAQAF------------------------
Q98I67          IGYAAGPVQLIKAMDMIQGQQTSGACTIAQWASVEALNGPQDF--IARNKAIFQGRRDLV
AATC_RHIME      MGFAVGNERLIAALTRVKSYLDYGAFTPIQVAATQALNGDGS--DIAEVRAIYKRRRDVM
AAT_STRGR       ------------------------------------------------------------
Q8YA73          VGFAVGNSEVIEAINLIQDHMYVSLFPGIQDAAIEALTGDQAC--VRELTARYENRRDAF
Q8ZI88          VGYLVADAQIVDVLFGCTVNVVGATNTSSQAGMIAALDEPSF---MGEYTQIFERRRKVV
O67864          VAFVVGNEILIKNLAHLKSYLDYGIFTPIQVASIIAL--ESPYEIVEKTAKVYQKRRDVL
Q93RH7          LAFAAGNADMIEALNLIQDHLFVSIFPAIQDAGVAALSDPRAKEAIAALNQRYDQRREAF
Q8R7H1          LGYIAAEHEFIEAMNKIHQYTTICAPITAQYAAIKGIYECEE--DIIKMRETYDQRRRFI
Q9A0S0          VGFAAGEPEIISAMSKIIGQTTSNLTTVSQYAAIEAFCGSQS--SLEEMRLAFEERLNIT
AAT_SULSO       LGYIVAKREIIQKMGILAANVYTAPTSFVQKAAVKAFD---TFDEVNQMVSLFKKRRDVM
Q8ZW57          LGVVIAPRDVAREVARFNKATINVPPTHAQRAIASVIDILP--KRREEVSAAYRRRAELA
Q982E0          IGYLAGPKEVIDAVKALQSHTTSNPNVVAQHAVLAHLQRSDGSYEAGLRAQLASARQVGL
Q982E3          LGYLAGPPEIVAAAKKLQSDMTSNANVIAQHAVLHHLEVGDCSFERKMHQRLSKARHTGL
O66630          IGMAVGNKELVAGLGKVKTNVDSGQFGAVQDAGIVALNLPEE--EVEKIRDVYRERKKIM
AAT2_METJA      IGYVISNDEIIEAILKLQQNLFISAPTISQYAALKAFEKETE-REINSMIKEFDRRRRLV
O29838          IGFACGNRDILAGLLKVKTNVDSGVFEAIQEAAIAAMDGPDR--VIEENCKVYQRRRDLL
Q9HQK2          LGWVVCPRRYVDTINAIAQNTLICAPSFVQAGAEAAIRHGTDW--LDGVRADYRTRRDIL
AAT1_BACSU      IGYAAGSEDIIKAMTNLASHSTSNPTSIAQYGAIAAYNGPSE--PLEEMREAFEHRLNTI
Q9YE99          LGYLYGPGWIVGKIRLVSEELVYCPPSIAQVAAKIYLEDEGRLRHLEYAREFLKTRMEAM
Q9R6Q3          IGLAVGDPEIIAAMTKIASQTTSNPTAVAQYAAIEAFEE--NDKSFEKMHAAFEERLNKI
Q97AE8          IGYAHTNSEILDRMKKLLRCTINGVNSATQYGAVAALTGPQEF--VGQMRSEYQKRRDII
O30304          IGFTIADEKLLDPMLKVHQVNGVCAPAFAQKAVAEVMAERLFDSIVSRMVKEFRQRRDYL
Q92QJ6          IGWMVLPEARVRAFERIAQSLYISPPELSQIAAEAALG---AHEELDGYKRAYAANRALL
AAT_PYRHO       LGFVVAPSQVIKEMTKLHAYVIGNVASFVQIAGIEALRSEESWKAVEEMKKEYNERRKIV
Q98KW9          MGWLTGPQDMITLLINLNLVTTYGLPAFISIACAEALENRYG---VKEIAERYAARRTVL
Q8VS38          IGYIAGRSDVVKAISTLQSQSTSNPNSIAQAAAVEALNGDHSF--LKERMGIFKSRRDFV
Q8XMH8          LGYVCAHKVLIDAMKKIHQYAIMCSPTTAQYAAIEALKNGDE--SVFEMAREYNRRRRVL
```

```
AAT_PYRAB       LGFVAAPSWIIEKMVKFQMYNATCPVTFIQYAAAKALRDERSWKAVEEMRKEYDRRRKLV
Q8VS39          IGYVAGRSDVVKAISILQSQSTSNPNSIAQAAAVEALN----------------------
Q9KE01          IGFAVGNPSVIKAIETLQDHYFCSLFGGIQAAAAHALLSDQS--NVTTLVQTYEERRNAL
O05237          LGFAAAPSLLRDAMLKIHQYAMMCAPAMAQFAALEGLKNGME--DVEKMKKSYRRRRNLF
AAT_AQUAE       IGYVACPEEYAKVIASLNSQSVSNVTTFAQYGALEALKNPKSKDFVNEMRNAFERRRDTA
Q8REF4          VGFLISKNKDFMTYIMKLCQARLAAPTVEQYAVASLMKAPKEY--FKEIKEIYKRRRDII
Q9PPF7          FGYMASKNALISAVKRLQGQSTSNICSITQHAAIPALNGECD-KDIEKMRQAFEKRRNLA
Q8Y1I0          LGWLVAPTELVPQFEKVAQNLFICASAVAQHAALACFEPE-ALAIYEGRKAEFHRRRDFI
Q8RR70          VGFLAGPVPLVKAATKIQGHSTSNVCTFAQYGAIAAYENSQDC--VQEMLAAFAERRRYM
Q9HRX4          LGWVAASERRAERMLRVHQYVQACASAPAQYAAEAALSGPQG--VVDEMVAAFEARRDVL
Q97M25          IGFIFAPEYLTSEIFKLHQYGSTCSCSISQYAALEALTN--GFNDSEYMKKEYIKRRDFI
Q9P9M8          LGYMSEVREAIDRLARIRL----CPNTPAQFAAIAGLTGPMDY--LKEYMKKLKERRDYI
AAT_THEAQ       IGYACGPKAVIKAMADVSSQSTTSPDTIAQWATLEALTNREASMAFIAMREAYRKRRDLL
Q8XXV2          VGYVAAPAALSAEFRKVHQFNVFTVNTPVQHGLAAYMADPRPYLELPAF---YQHKRDLF
Q8TS80          LGFAMGAPDIIHSMMMIHQYSMLCAPITAQVGAIEALRNGKE--EMERMVREYDRRRRFI
Q8VXZ8          IGWAIAPPHLTWGVRQAHSYLTFATSTPAQWAAVAALKAPESY--FKELKRDYNVKKETL
Q8XRB5          LGWVAGPRDLIQALDTFLSQSAGNCCSISQAAAAAALNGDQRF--VAESVAVYKQRRDTT


1gdeA           WKRLNEMGLPTVKPKGAFYIFPRIRDTGLTSKKFSELMLKEARVAVVPGSAFGKAGEGYV
Q8Y0E8          LSRLGGAGLRTPRPQGAFYLFPDAQAFAIDDDQALAAYLLDSGVAVVPGSGFGM--PGCL
Q9BVY5          VRLLESVGLKPIVPDGGYFIIADVSLLDPDLSKFVKWMTKHKKLSAIPVSAFCNSEEKFV
Q8RAK7          VERISKMGLKAVRPQGAFYVFVNIEEYVKGSLDFATLLIEEANVAVVPALPFGM--DNYI
AAB2_RHIME      VNGLNAIGLDCRVPEGAFYTFSGCAGVAESDTDFCAYLLEDSHVAVVPGSAFGL--SPYF
Q55453          TLGQLSDYCRLVVPQGAFYCLLEV-NSPLTDLELVKRLIDEFKVAVLPGSTFGVDSGCYL
AATB_RHIME      VNGLNAIGLDCRVPEGAFYTFSGCAGVAESDTDFCAYLLEDSHVAVVPGSAFGL--SPYF
Q8YTF2          IQGLGELGWDVPKTKATMYLWVKCP-VGMGSTDFALNLLQQTGVVVTPGNAFGVAGEGYV
Q9KC79          YDKLVAIGVSCVKPKGAFYLFPNVEEAAATVDDFVKVLLEEEKVALVPGSGFGA--PDNV
Q92S71          DALIATNRVETLKPDGALYAFLKI-DGVTDSRAAALDIVDRTGVGLAPGTAFGEGGALFM
Q93703          SILADVPGMRVVKPKGAMYMMVNISRTAYGSDSFCQNLIREESVFCLPGQAFSA--PGYF
Q8YMS6          LDRLNAIGLSTAKPDGAFYLFPDISKTGLKSLEFCDALIEEHKVAVIPGIAFGA--DDNI
Q9HV76          LPALRELGFGIVEPEGAFYLYADISAFGGDAYAFCKHFIETEHVAFTPGLDFGRHQAGHV
Q8U3E6          LQRLRKMPYIKVEPKGAFYVFPNISETEMSSEEFSEWLLEKAKVVVIPGTAFGENGEGYV
Q9HUI9          IECLADPGLRPLRPDGGMFVMVDIRPTGLSAQAFADRLLDRHGVSVLAGEAFGPSAAGHI
Q8ZVJ5          VAALRKYVPEAIVPAGSMFIFVDLSRYISDGESFARELLERYGVAVVPGSYFSEIYRAAV
AAT_METEX       ------------------------------------------------------------
Q9ZLG5          HAKINAIGLNALKPDGAFYLFINIGSLGGDSMRFCHELLEKEGVALVPGKAFGL--EGYV
AAT1_METJA      YNGLKDI-FKVNKPDGAFYIFPDVSEYG--DGVEVAKKLIENKVLCVPGVAFGENGANYI
Q8XJ54          YKRLKSMGLEVEKPKGAFYIYPNISKFGLTSEEFCHRLLKEGKVACVPGDAFGKGGEGYI
Q8UFR3          MARLPEIGLPLASPDGAFYAYVDTSRFSNDSMDFAKRMLAEIDVAATPGMDFDPEGHRAL
Q9W6U2          ASLLKSVGLKPIMPEGGYFMTADFSSIKVDCDRFVKWLIKEKGLATIPVSAFGKEFDKYI
Q9K7L1          VKTFTEIGLTCPMPGGAFYAFPSVKETGLTSEEFAERLLMEEHVAVVPGNVFGEGGEGHI
O66737          YGELKKLFKIDAKPQGAFYVWANISDYSTDSYEFALKLLREARVAVTPGVDFGKNKKEYI
Q9Y9P0          YDILSQEGIEPYLPEGAFYMFPRVRKTGLSVEQLAEKLLYSYGVLVLPGTSFPESGREHV
Q99V44          VSELTKLGFETAQPEGAFYIFPSIKHITDDDFEFCVDLLESTHLAIVPGSSFTEFGKGFV
Q8RA61          VERVNKIDLKCSTPKGAFYVMVNIEKTINGSLDFASALIDGANVAVVPALPFGM--DNYI
PATA_BACSU      YDRLVSMGLDVVKPSGAFYIFPSIKSFGMTSFDFSMALLEDAGVALVPGSSFSTYGEGYV
Q98NB8          MKRLPELGFPLAAPDGAFYAFCDVTRHSNDSMAFARKMLAEAHVAATPGRDFDTAGHRTM
Q97PQ9          YPLLAEVGFEVVKPQGAFYLFPNVKKAMTDVTDFTTVILEEAEVALVTGAGFGA--PENV
Q9KAU1          VQVLEDAGFEFVYPKGGFFIFPKSPIADDT--KFCQVAAEEFQLLVVPGIGFGRA--GHF
Q9V0G5          YKRLNEIGISTTKPQGAFYIFPRIEEGPWDDKEFVLDVLHNAHVLFVHGSGFGEYGRGHF
AAT_PYRKO       SKMLR--GFDFVEPEGAFYIFLRTPQDGMA---FAERLLSR-GVAVFPGMAFGDY-PNFI
Q9L0L5          VRMLNEIGVLCPEPEGAFYAYPSVKALPQDSVELAALILEEAEVAVVPGEAFGT--PGYL
Q979X6          LSILDSDKLNVTEPDGAFYVFPEY-NSDISSNKVSEDLLEKYQVVVTPGSAFGRQGEHHF
```

```
Q8RCV4          LESVREMGLSCFEPKGAFYIFPSIKTTGLTSMEFAKRLLYEAKVAVVPGNAFGEHGEGYV
AAT_SYNY3       VEGINQAGLSCPNPKGAFYVFVDIAKTGLNSLEFSARLLESHQVAVIPGAAFGA--DDCV
Q8XJT3          IELISEIEDLTIEPKGAFYVMIDVSKVLKGSMEFANLLLKEENVVVIPGIAFGE--DNFI
Q97KB8          YSGLSSIGVICSKPEGAFYILAKLPVD--DSDAFAKWLLTDETVMFAPASGFYATGKSEA
Q8ZDK4          WELINQIGVSCVKPQGALYMFPRIDQKRFNLKKLVLDLLLQEKVLLVQGSAFNWPYPDHV
O59096          WKRLNEMGLPTVKPKGAFYIFPRIRDTGLTSKKFSELMLKEARVAVVPGSAFGKAGEGYV
Q98AR6          DVLAWLTRVPVPRAQGGFYFYLDLSHLSTTADDIVTALLAETGVAAVSGAAFGD--PAGL
Q8TT07          LKRLLGMGLEVREPMGAFYVLADARKFGNDSLELSRSILNEAGVAVTPGVDFGNGAEGYL
Q9R096          IRSLQSVGLKLWISQGSYFLIADISDDEPYDRRFAKWMIKNMGLVGIPVSTFFSRPDHYI
Q9RNK6          VDGLNAIEGISCPKPQVPSMSIRAFPSLLARKPHQVKSLKKTWIS----------------
Q9SIE1          VKSLGDIGVKISEPQGAFYLFIDFSALINDSSSLALYFLDKFQVAMVPGDAFGD--DSCI
YD91_METJA      VKILNEVGFKARMPGGTFYLYVKSPTKAKTAEDFSQYLIKEKLISTVPWDDAGHDENGNP
Q9CPI6          YELINQIGVSCIKPQGAMYMFPKIDIKKFDDEKMVFDLLAQEKVLLVHGRGFNWHSPDHF
Q8U097          SKILR--GLEFYEPDGAFYIFLKTPIDGLEF----VYKLLERGVSAFPGIAFGNY-QNFI
Q9CEK7          YGQLSEVGFELVKPNGAFYLFPKVTKAMSDVTDFTTAILEEAGVALVTGAGFGS--PENV
AAT_BACSP       YPKLSAIGFKVVKPQGAFYLLPDVSEAAQKTGEFASALLTEANVAVIPGSGFGA--PSTI
Q9I015          WELLNDIGVSCVKPMGALYAFPRIDPKVHNDEKFVLDLLLSEKLLIVQGTAFNWPWPDHF
Q9KQM1          WELINQIGISCVKPKGAMYLFPKIDTKMYPIKKMVLDFLVQEKVLLVQGSGFNWPKPDHF
Q9A8H2          IKSMKAAGWDIPNPPASMFAWAKIPEAEAGSMLFSRLLIEEAGVAVAPGIGFGEYGEGYV
O27916          FRSLTDMGLECVLPGGAFYMFPYAGD----SEEFTKLSL-EAGVAVVPGSAFGNEGKGYI
AAT_BACST       YDKLVQIGFTCVKPQGAFYLFPNAREAARTVDEFVAALLEEAKVALVPGSGFGA--PDNV
AAB1_RHIME      VNGLNAIGLDCRVPEGAFYTFSGCAGVLKTDTDFCAYLLEDAHVAVVPGSAFGL--SPFF
AATA_RHIME      VSMLNQKGISCPTPEGAFYVYPSCAGLIETDEDFVSELLETEGVAVVHGSAFGL--GPNF
Q9HHD3          YKRLNEMGISTQKPQGAFYIFPKIEEGPWKSKEFVLDVLHNAHVLFVHGSGFGEGGEMHF
Q8Y525          AKRVEKIPYLSLHPKGSIYAFINISKTNMDSVSFTEYVLKETQVLVIPGLAFGESGDNYV
AAT_RICPR       LSILEEVTYFEYKPEGAFYLFVKCDKIFGTSNNFSEYLLEEAKVAVVPGIAFGL--DGYF
Q8YIC8          VSMLNQTGLQCPTPEGAFYVYPSCAGLIETDKDFVTELLETEGVAVVHGSAFGL--GPNF
Q9HK41          YEAVSESRLEPIKPHGTFYLWNRIKEYPSDSWGMTSYLLEKTGVGSSPGPVFGPAGEGYI
AAT_RHILP       VSMLNQKGIVCPVPEGAFYVYPSCKGLIETDEDFVSELLESEGVAVVHGSAFGL--GPNF
Q53951          ------------------------------------------------------------
Q9HKR7          YGILSETDMKVHKPEGAFYMFPGYSKD-IPSEKIAEMLLNQEHVVVTPGSAFGDRGRHHF
O25383          HAKINAIGLNALKPDGAFYLFIHIGSLGGDSMRFCHELLEKEGVALVPGKAFGL--EGYV
Q9CAP1          VKGLKEVGFTVFPSSGTYFVVADHTPFGMENDAFCEYLIEEVGVVAIPTSVFYLNGKNLV
AAT_THETH       LEGLTALGLKAVRPSGAFYVLMDTSPIAPDEVRAAERLL-EAGVAVVPGTDFAAF--GHV
Q9RWP3          VAGLNALGLPTPTPQGAFYVMADTRAIHTDELEAARIILDEAQVAVVPGTDFAAPGQ--V
Q9UZ63          LKRLKEMGIRVKEPKGAFYVFPSIKDTGMSSEKFSEWLLEKARVVVIPGTAFGKMGEGYV
Q8U1F5          WKRLNEMGLPTVEPKGAFYIFPRIKDTGLSSKEFSELMLMEAKVAVVPGSAFGRAGEGYI
Q08415          IRSLQSVGLKLWISQGSYFLIADISDDEPYDRRFAKWMIKNMGLVGIPVSTFFSRPDHYI
AAT2_BACSU      TACEKELGWKMEKPKGTFYVWAEIPNTFETSHQFSDYLLEHAHVVVTPGEIFGSNGKRHV
Q9X8S5          AAGLAEAGFGVYRPAGTYFVTTDIRPLGRDGFAFCRSLPERAGVVAVPNAVFYNHGAPFV
Q939K8          YEKMKNLGFEIARPNGAFYIFAKIPDGTQNSMNFCVDLAEKNKLAIIPGSAFGAAGEGFV
Q9RAT0          IEKMSALGFKIIEPDGAFYIFAKIPDLEQDSFKFAVDFAKENAVAIIPGIAFGQYGEGFV
Q9CJE0          IEKMSDLGFKIIEPDGAFYIFAKIPDLEQDSFKFAVDFAKENAVAIIPGIAFGQYGEGFV
Q972A2          YEELKKIGIQVHKSQGAFYMFPFIGEINLSVKDFSLKLIEEKGVTTIPGEVFPLEGKDFV
Q9XBE6          LADLARIPGLTTVPDGGFYAFVGIRVAESTSAAVAEYLLHECGVAVVPGSVFGRAGEGFV
Q9HRM6          VDLFAEHGTDVSTPDGAFYLMLPVADD---DQAWCQDALKDAHVATVPGSAFGA--PGYA
Q92JE7          LSILKRVKYFEYKPEGAFYLFVKCDKIIANSNDFAEYLLEEAKVAVVPGIAFGL--EGYF
Q9AA68          TAGLRDAGYVVLESQGTYFLNVDLAASGLDDVTFCERCVTEHGVAAIPVSAFFAEDTTVV
Q8U821          LGQLAEIGLRCHKPEGAFYIYPNISGLIESDVDFVMALVDEHHVATVQGAAYGM--SPYF
Q92D16          QDRLEKMGFTVIPPDGAFYFFVKLPDDITNSFDWAVRLAEEAKVAVVPGNAFSEKGDRYF
O28650          VKRLSEI-FEIKKPEGAFYAFPKISSTGMSSEEFAEKLLLEKSVAVVPGNAFGECGEGYV
AAT_THEMA       VERLKKMGVKFVEPEGAFYLFFKVRGD---DVKFCERLLEEKKVALVPGSAFLK--PGFV
Q8YP73          VKGLQSLGWPIEPPKATLYVWVPVPP-GYTSTEFTTLLLDKCGIVVPPGVGYGASGEGYF
O53870          AAGLTEIGFAVHDSYGTYFLCADPRPLGYDSTEFCAALPEKVGVAAIPMSAFCDPAAGLV
```

```
Q8TR00          VKGLNELGMECAFPKGAFYAFPKVENSAEVASKMI-----SNGVVVVPGTAFGSEGDGYI
Q97I35          SKRVNNTGIHCLLPKGAFYVMMNISNLINNSVDFSKELLSENKVAVVPGTGFGN--DNYV
Q8XGH1          WELINDIGVSCVKPRGALYMFPRIDAKRFDDQKMVLDFLLQEKVLLVQGTAFNWPWPDHF
Q8UDD1          VSMLNQKGISCPSPEGAFYVYPSCAGLIETDTDFVSELLEAEGVAVVQGSAFGL--GPNF
Q54188          ------------------------------------------------------------
Q8Y606          YPELNSIGFKPKKPDGAFYFFIEVKEAAQDVDAFVAALLEEAKVAVIPGSGFGM--PDYI
Q8TPT6          IDGLNKIGIECKKPDGAFYAFANVSEYG-NGTEVAERLLKEAHVAVTPGIAFGASGEDFI
Q98H83          VEGLNALPNISITPKGAFYAFPNVSKTGWKAKKLASALLDDAGVALIGGPDFGILGEGYV
O31665          ITACREIGWDVTAPAGSFFAWLPVPE-GYTSEQFSDLLLEKANVAVAAGNGFGEYGEGYV
O28151          MAKLDEMGIEYAPPKGAFYIFMNV---GRDSNEFCEEFLKREYVALTPGSAFGVAYKSWV
Q8Y8A4          QDRLEKMGFTVIPPDGAFYFFVKLPDEAENSFDWAVKLAEEAKVAVVPGNAFSEKGDRYF
Q92AB1          YPELSSIGFKPKKPDGAFYFFIEVKEAAQDVDAFVAALLEEAKVAVIPGSGFGM--PDYI
O54170          VGRLSGVGLHLVPPEGAFYGFLRYDAD--RPSEMVARELAACQVLVRAGAEYGPSGEGHL
Q9A830          VSMLNQTGLHCPTPEGAFYVYPSCAGLIESDEDFATELLESEGVAVVHGAAFGL--SPFF
Q8XRN7          TALSALPGVDVRVPEGAMYAFFRIPGA-QDSLALCKQLVREARLGLAPGSAFGPEGEGFV
Q8W360          ------------------------------------------------------------
Q929C3          AKRVAEIPYLSLHPKGSIYAFINISKTKMDSVSFTEYVLKETQVLVIPGLAFGESGDNYV
O86587          WEKLNEIGISCVKPKGALYAFPRIDPAVHDDERFVLDLLLREKIQVVQGTGFNWPSPDHF
O59044          YKRLNEIGISTTKPQGAFYIFPRIEEGPWKSKEFVLDVLHNAHVLFVHGSGFGEYGKGHF
Q943I5          VKSFKELGVKISEPQGAFYLFIDFSSTIKDSESLCMFLLEKAQVALVPGDAFGD--DKCI
Q97TA8          IEKMTALGFEIIKPDGAFYIFAKIPAGNQDSFAFLKDFAQKKAVAFIPGAAFGRYGEGYV
AAT_STRVG       VKMLNEIGVFCPTPEGAFYAYPSVKELPQSSVELAALILDEVEVAVVPGEAFGT--PGYL
Q8RQG1          ----------------SFFAWLPVPK-GYTSEQFSDILLEKAHVAVAPGVGFGKHGEGYV
Q8XQJ3          VGMINAVGMRCETPAGAFYAFASCDGLIRSDEDVSNALLDERGIGVVPGSAFGL--GPYL
Q8TQ40          IEGLTAMGLEVKPPKATFYVWAPVP-TGFTSIEFAKLLLEEAGIVATPGVGFGDAGEGYV
AAT_RICCN       LSILKRVKYFEYKPEGAFYLFVKCDKIIANSNDFAEYLLEEAKVAVVPGIAFGL--EGYF
Q9RWJ7          SSGLRSLGAQVHEPQGTYFLTAQHPTWT------AEGLVESGAVAVIPGEAFYVTPAGLL
Q97GI7          YKRLKDMGFEVRLPKGAFYIFPDISRFGMTSEQFCEKLLNEAKVAIVPGSAFGEKGEGFA
O95335          ------------------------------------------------------------
Q98I67          VSMLNQRGITCPSPEGAFYVYPSCAQLIDTDEAFCSELLEAEGVAVVFGSAFGL--GPNF
AATC_RHIME      VESFGKAGFEVPPPPATMFAWAKIPEKHLGSLEFSKLLVEKADVAVAPGIGFGEQGDDYV
AAT_STRGR       ------------------------------------------------------------
Q8YA73          ISACREIGWEAVAPAGSFFAWMPVPED-FTSSEFADYLLEEVSVAVADGSGFGEFGEGYV
Q8ZI88          FEMMNAIGVCMVMPESGFLSWIDISKLG-TSTFICDYLLQHAHVMVNSGVPYGQGGEGYI
O67864          VEGLNRLGWKVKKPKATMFVWAKIPEWNMNSLDFSLFLLKEAKVAVSPGVGFGQYGEGYV
Q93RH7          VQAAAQIGWQAFPSKGSFYAWMPVPE-GFNSQNFADLLLEEAHVAVAPGIGFGQEGDSYV
Q8R7H1          VNGFREIGLDCFEPKGAFYIFPSIKKTGMTSEEFCEKLLKEEKVAVVPGNAFGPSGEGYV
Q9A0S0          YPLLCQVGFEVVKPQGAFYFFPNVKKAMEMTGSFANAILEEVGLAVVSGAGFGA--PENV
AAT_SULSO       YDELTKVGVEVSKPNGAFYMFPNVSKILKTSKSLAIKLIEEKGVVTIPGEVFPLNGKEFL
Q8ZW57          ERLLR---LPFVKPGGAFYLFPRVSEGCFDK--------ALAAGVSVLPGELYGR--GGHV
Q982E0          SMLSSLTQVPIPRAQGGFYFYLDISKLVWSAREIVSALLTDAGVGVVSGSVFGD--PAGL
Q982E3          RILAGLEGVTVPRADGTFFFYLDLKRLVRSADDIARLLLDETDVATVAGGFFGDMNG--L
O66630          TEALEKIGLEIYRSDYTFYLWIKVPE-GYTSAEFVGRLIDEAGIVCTPGNGFGEYGEGYF
AAT2_METJA      LKYVKDFGWEVNNPIGAYYVFPNIGEDG---REFAYKLLKEKFVALTPGIGFGSKGKNYI
O29838          VEGLRDVGIDAEKPKATFYVWAKV---GGSSIEFVKQLIDKAGIVATPGIGFGKSGEGFV
Q9HQK2          CAAAERWGFDPYTPAGAYYMLLDVSRLG-AAPAVADALLETAGVAVTPGPDFGANATEYV
AAT1_BACSU      YAKLIEIGFSCVKPEGAFYLFPNAKEAAQSCGEFVKALLEEEKVAIVPGSGFGS--PENV
Q9YE99          AGALEEMLPEAVRPGGSMFITVDLSSASITSEDLSVKLLDEESVATVPGRYFGPSGDTML
Q9R6Q3          YLQLSEVGFELVKPNGAFYLFPKVTKAMTDVTDFTTAILEEAGVALVTGAGFGS--PENV
Q97AE8          FNAVSKSRLEPVKPGGTFYLWAKIKEEVKDSWDMTYYLLDKTGVGSSPGPVFGPAGDGYI
O30304          YSELSKI-YDVVKPEGAFYMFVNVNQDCMEY----AENLINLGVAVTPGLPFGDGNETYV
Q92QJ6          LERLPQIGFSIASPDGAFYAYADVSRFTNDSMAFARRMLAEIDVAATPGFDFDPEGHRTM
AAT_PYRHO       VKRLKNMGIKVKEPKGAFYVFPNISGTGMSSEKFSEWLLEKARVVVIPGTAFGRMGEGYV
Q98KW9          LDAVRGMNNVSVGSEGGMYVMLDISDVEPDDEKFAWALLDQEKVGVMPGSSFGEAAAGHI
```

```
Q8VS38          IEKLNSPGLLASIPQGAFYLFVSCKDLLGKSTKFAEYLLEDHLVAVIPGIAFGL--KNFI
Q8XMH8          VDGFRSMGLDCFEPLGALYVFPCIKSTGMSSDEFCEKLLLEEKVLAVPGNAFGECGEGFI
AAT_PYRAB       WKRLNEMGLPTVKPKGAFYIFPRIKDTGLTSKEFSELMLMEAKVAVVPGSAFGKAGEGYV
Q8VS39          ------------------------------------------------------------
Q9KE01          VKAARAIGWDVEAPKGSFFAWFPVP-SGFTSEEFATYLLKKARVVVAPGKGFGEHGEGYV
O05237          VESLNEIGLSCHHPGGRFLCFSIYQKHGNEFKSLPEELLTQEKVAVV-------------
AAT_AQUAE       VEELSKIGMDVVKPEGAFYIFPDFSAYAGGDVKLSEFLLEKAKVAVVPGSAFGA--PGFL
Q8REF4          VNSLNKIGVTCSTPKGAIYAFAKLPVE--SSEDFCKWLLTESTVMLAPGEGFYETGKNEV
Q9PPF7          LDILKQIPNISVKPEGAFYLFVNIQKIEKDSMKFCQKLLEQEKVAVVPGVGFGM--DGYF
Q8Y1I0          VPALESLGFQVVKPDGAFYVYADCRGVNHDADALTQSMLNDAGVVLVPGLDFGPTAHHYI
Q8RR70          LDALNAMGLECPKPDGAFYMFPSIAKTGRSSLDFCSELLDQHQVATVPGAAFGA--DDCI
Q9HRX4          LDGLEAMGLDVPTPGGAFYAMPTVPDGWIDE-------VVDRGVVVVPGSAFGDHGAGTA
Q97M25          YKELVSMDFDVVKPEGAFYIFPSIKKFNMTSLNFSLKLLEKEHLAVVPGSAFSHYGEGYI
Q9P9M8          YKRLNEIGISTTKPQGAFYIFPKIEVGPWKNKEFVLDVLHNAHVLFVHGSGFGEYGAGHF
AAT_THEAQ       LEGLSRIGLEAVRPSGAFYVLMDTSPFAPNEVEAAERLLM-AGVAVVPGTEFAAF--GHV
Q8XXV2          RAGLEHTRFKLLPCQGTYFQCVDYSASDLPEAAFAKWLTSEIGVAAIPVSAFYSQESGVV
Q8TS80          VKGFNSIGLECGNPKGAFYAFPYIGGTGLSSSDFAERLLEEKKVVTIPGDVFGEAGEGFL
Q8VXZ8          VKGLKEVGFTVFPSSGTYFVVADHTPFGMENDAFCEYLIEEVGVVAIPTSVFYLNGKNLV
Q8XRB5          LARLNAIGLTCRSPDGAFYLYVNCAGTLDTDTDVVMYLLEREGVAIVAGTAYGL--SPYF


1gdeA           RISYATAYEKLEEAMDRMERVLKERKLV
Q8Y0E8          RLSYATSEARLELAATRLAGALR-----
Q9BVY5          RFCFIKKDSTLDAAEEIIK---------
Q8RAK7          RISYATSMENIEKGLDRIENFLNK----
AAB2_RHIME      RISYATSEAELKEALERIS---------
Q55453          RIAYGALRQTASVAIARLEQGLKS----
AATB_RHIME      RISYATSEAELKEALERIS---------
Q8YTF2          RISLIADCDRLGEALDRIKQ--------
Q9KC79          RLSYATSLELLEAAIDRIFRFVEAKR--
Q92S71          RACFLRDPLQIAEAADRLQRYILSR---
Q93703          RVVLTCGSEDMEEAALRI----------
Q8YMS6          RLSYATDLATIEKGLDRLEKFVRSR---
Q9HV76          RFAYTQNLPRLQEAVERIARGLKNWR--
Q8U3E6          RISYATSREKLIEAMDRIEKALEE----
Q9HUI9          RLGLVLGAEPLREACRRIALCAAE----
Q8ZVJ5          RISFVTEPPRLEEGIRRIGEALKA----
AAT_METEX       ----------------------------
Q9ZLG5          RLSFACSEEQIEKGIERIARFVKSK---
AAT1_METJA      RFSYATKYEDIEKALGIIKEIFE-----
Q8XJ54          RISYCYSKDELERALDKLEAFVKTLK--
Q8UFR3          RISYAGSVSDIAEAVGRIAGWLK-----
Q9W6U2          RFCFVKEEATLDAAAEILKKWSQEQ---
Q9K7L1          RCSYATSMEHLETALERIGRFVQKCK--
O66737          RFAYTRKIEELKEGVERIKKFLEK----
Q9Y9P0          RLSFATATSDVKEGAEIIVRASRE----
Q99V44          RISYAYEMDVLKEGMKRLAKYLNTK---
Q8RA61          RMSYATSLENIKKGLDRIEDFLSKS---
PATA_BACSU      RLSFACSMDTLREGLDRLELFVLKKR--
Q98NB8          RFSYAGSHDDMVEAMARIERWLR-----
Q97PQ9          RLSYATDLDTLKEAVERLKAFMGSEN--
Q9KAU1          RLSYSVSMEQIERSAPVLKKL-------
Q9V0G5          RAVFLPPVEVLEEAMNRFEKFMREK---
AAT_PYRKO       RISLSG------KGLERGLRVIRE----
```

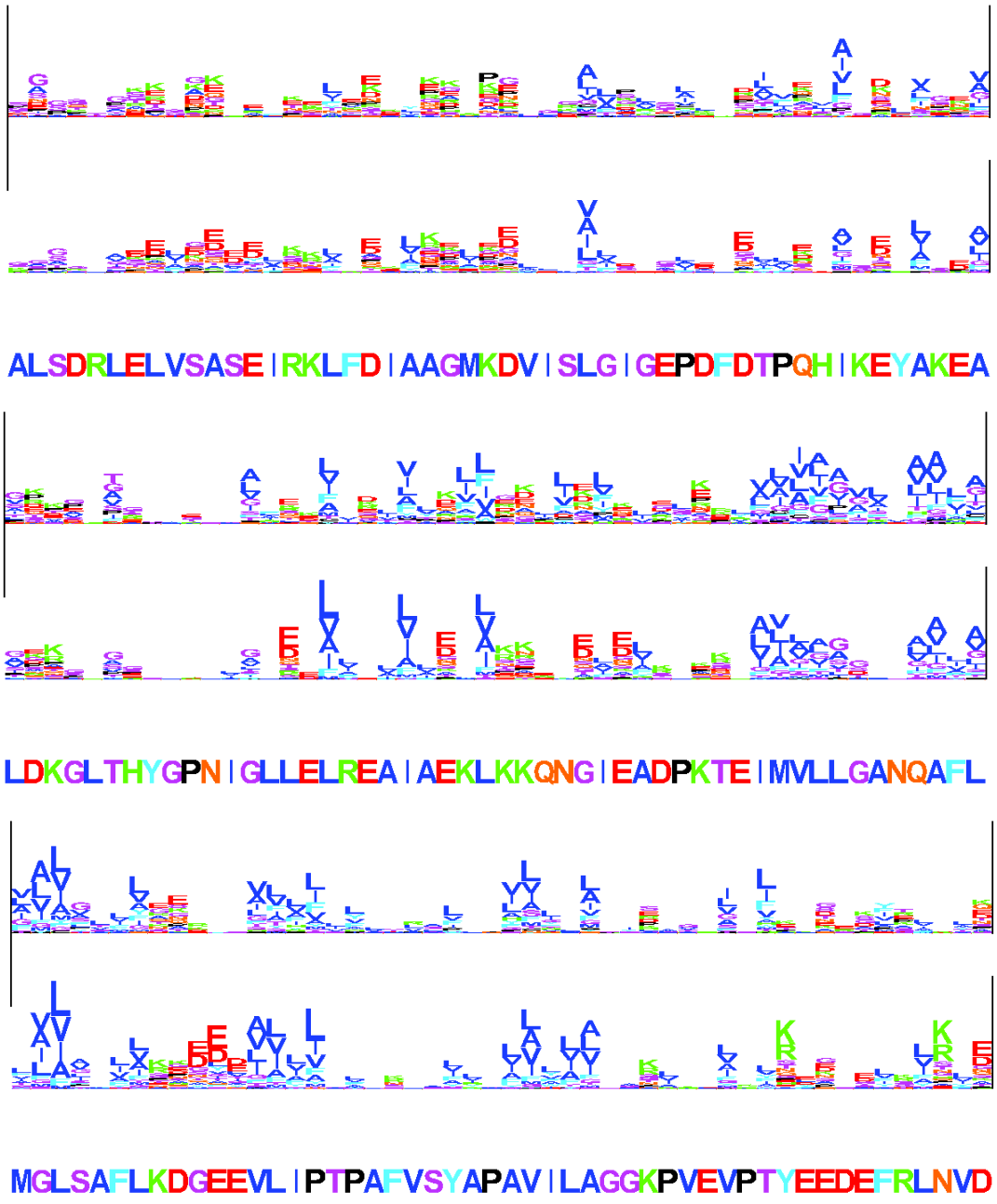
```
Q9L0L5          RLSYALGDEDLVEGVSRIQKLLAEAR--
Q979X6          RISFATSEDIIKEGAERIIKFFAS----
Q8RCV4          RLAYATSMENLEEAVKRMKEFMAK----
AAT_SYNY3       RFSYATDMDTIKQGLAELERFVST----
Q8XJT3          RLSYATSKEEIIKGLKRIKEFVNK----
Q97KB8          RFSYCTSIEDIENSIVILKKALEEYN--
Q8ZDK4          RIVTLPRVDELEMAVRKLGRFLETYH--
O59096          RISYATAYEKLEEAMDRMERVLKERKLV
Q98AR6          RLSYGIPSEKLATGLARLVEFLNSWK--
Q8TT07          RFSYANSLENIAEGMDRLEAFLEK----
Q9R096          RFCFVKDKATLQAMDERLRK--------
Q9RNK6          ----------------------------
Q9SIE1          RISYATSLDVLQAAVEKIRKALEPLR--
YD91_METJA      TTEEKYEDMVLEEFKRRLEGMDLE----
Q9CPI6          RIVTLPHVHQIEEALTKLARFL------
Q8U097          RISLTS--DKLEKGLNIIKEV-------
Q9CEK7          RLSYATSLKNLEVAVARLKDWMNE----
AAT_BACSP       RISYATSLNLIEEAIERIDRFVK-----
Q9I015          RVVTLPRVDDLEQAILRIGSFLKGYQ--
Q9KQM1          RIVTLPHVEDLEIAISRFERFITT----
Q9A8H2          RIGLVENEHRIRQAARNVKKFIANADSI
O27916          RMSYATSYELIEEAMERLKTV-------
AAT_BACST       RLSYATSLDALETAVERIHRFMEAR---
AAB1_RHIME      RISYATSEAELKEALERIA---------
AATA_RHIME      RISYATSEALLEEACRRIQR--------
Q9HHD3          RSIFLAPVPVLEEAMDNLEKFMKER---
Q8Y525          RLAATQDISVLEEAFNRLAKL-------
AAT_RICPR       RISYATSMQELEEACIRIKH--------
Q8YIC8          RISYATSDELLEKACIRIQR--------
Q9HK41          RFAFSAATDHIQEAADVL----------
AAT_RHILP       RISYATSEEQLEEACRRIQRFCGACK--
Q53951          ----------------------------
Q9HKR7          RISFATSEDIIKEGLERLVKFMHT----
O25383          RLSFACSEEQIEKGIERIARFVKSK---
Q9CAP1          RFAFCKDEETLRGAIERMKQKLKRK---
AAT_THETH       RLSYATSEENLRKALERFARVL------
Q9RWP3          RLSYATSMDNIEEVLRRLEGVVR-----
Q9UZ63          RISYATSREKLMEAMDRMEKALSE----
Q8U1F5          RISYATAYEKLEEAMNRMEKVLKEKKLI
Q08415          RFCFVKDKATLQAMDERLRK--------
AAT2_BACSU      RISMVSKQEDLREFVTRIQKL-------
Q9X8S5          RFAFCKRLPVLEEAVGRLKTL-------
Q939K8          RLSYAASMEKLELAMERLTAYMATNK--
Q9RAT0          RLSYAASMDVIEQAMARLTDYVTKKR--
Q9CJE0          RLSYAASMDMIEQAMARLTDYVTKKR--
Q972A2          RLSFAVKEDDIREGIKRMKEFI------
Q9XBE6          RMSFAGSAEQLDRAVKRLA---------
Q9HRM6          RLSYAASTDRLEAAVDRLA---------
Q92JE7          RISYATSMEELEKACIRIEKTI------
Q9AA68          RLCFAKADATLDEAVRRLA---------
Q8U821          RLSYATSMEMLGEGCARIAQFCRDIR--
Q92D16          RLSYATSVNNLAEALDRMAQFLAK----
O28650          RLAYAVKFEKLKEAMDRIKEFVEEH---
AAT_THEMA       RLSFATSIERLTEALDRIEDFLNSR---
```

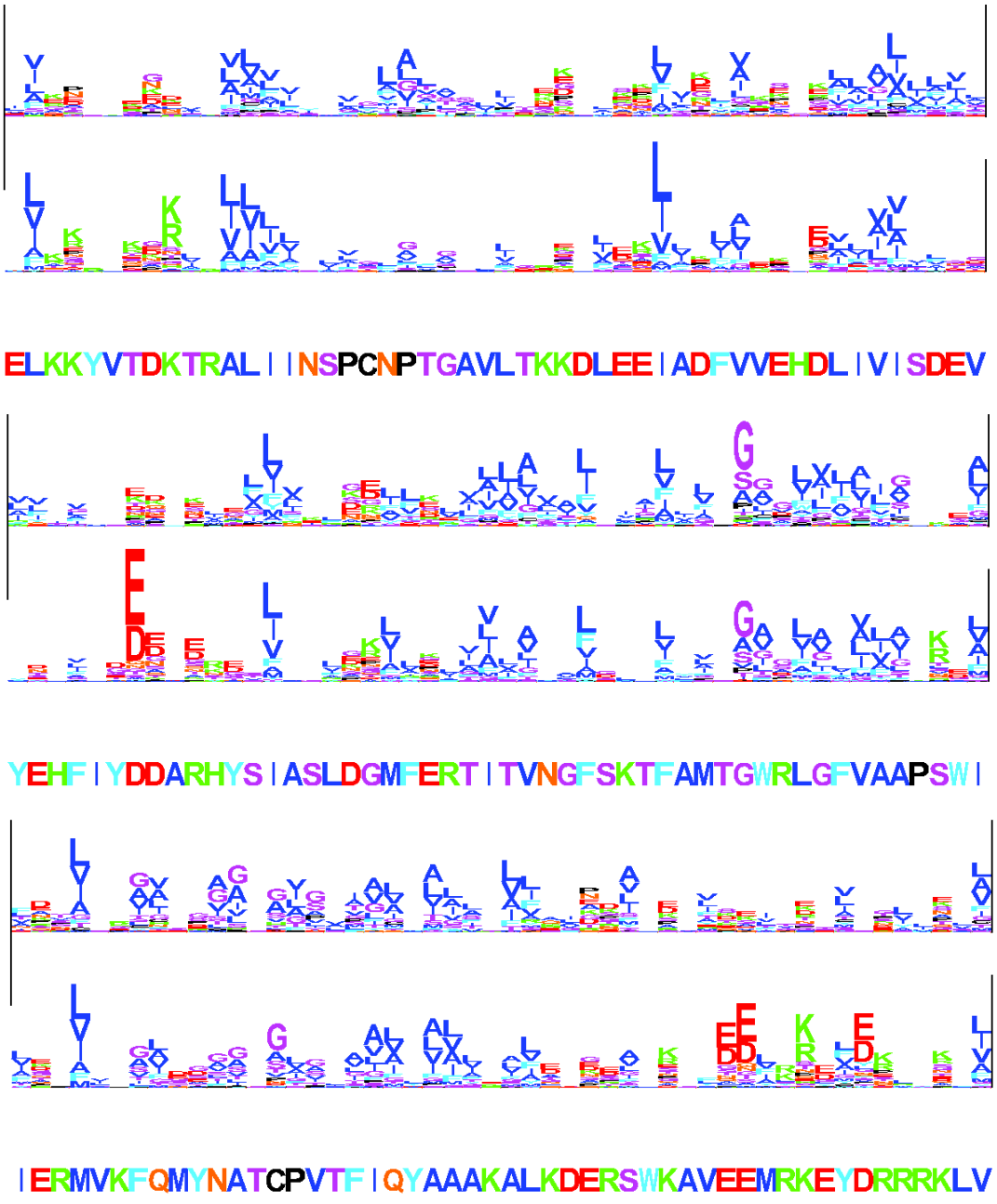
```
Q8YP73          RVALTISDERLHEAIQRMQ---------
O53870          RFTFCKRDDTLDEAIRRLSVL-------
Q8TR00          RISYAASMKDIEKSLAIMEKVL------
Q97I35          RLSYATSMDNIVKGLDEIENFIGKLR--
Q8XGH1          RIVTLPREDDLEMAINRFGRFLSGYH--
Q8UDD1          RISYATSETLLEEACKRIQR--------
Q54188          ----------------------------
Q8Y606          RLSYATNPDLFQEAINRIKSFMK-----
Q8TPT6          RISYATSIDRIREALERLEKIF------
Q98H83          RLSYANSEENILRALERIGAFLSK----
O31665          RVGLLTSEERLKEAAYRIGKL-------
O28151          RLSYATSRERIGEFLSRLERFL------
Q8Y8A4          RLSYATSFNNLAEALDRMAQFVEK----
Q92AB1          RLSYATNPDLFQEAINRIKSFMK-----
O54170          RISFAASEDDLWVGLERIVRYFAEAR--
Q9A830          RISYATSNEVLEDACSRIQRFCASVK--
Q8XRN7          RWCYACDVARLAAGVERLREFLR-----
Q8W360          ----------------------------
Q929C3          RLAATQDISVLEEAFNRLAKL-------
O86587          RILTLPHAEDLEAAIGRIGRFLSGYR--
O59044          RIVFLPPIEILEEAMDRFEKFMRER---
Q943I5          RMSYAAALSTLQTAMEKIKEAVALIK--
Q97TA8          RLSYAASMETIKEAMKRLEEYMRE----
AAT_STRVG       RLSYALGDEDLVEGVSRIQKLLAEAK--
Q8RQG1          RVGLLHAEDRLREAINRIDKL-------
Q8XQJ3          RIAYAIDDTALRTACTAIDTFARSLR--
Q8TQ40          RFALTKPVERIKEAVERMKKL-------
AAT_RICCN       RISYATSMEELEKACIRIEKTI------
Q9RWJ7          RFAFCKSRAELEQALERLARL-------
Q97GI7          RISYAYSKKELEECMNRIEKWVEVQNNL
O95335          ----------------------------
Q98I67          RISYATSDALLEEACTRIQRFTAS----
AATC_RHIME      RLALVENEHRIRQAARNIKRFLSS----
AAT_STRGR       ----------------------------
Q8YA73          RVGLLMDEERLEEAVTRVSKLHLFDKV-
Q8ZI88          RLVHGCDDEKLYAVLTRIQQALMQ----
O67864          RFALVENEHRIRQAIRGIRKAFRKLQ--
Q93RH7          RIGLLVEPDHLAQAVERIGQL-------
Q8R7H1          RVSYAYSIDKIAKALERIKRFAEK----
Q9A0S0          RLSYATDIETLKEAVRRLHVFMESNEI-
AAT_SULSO       RLSFAVNEEVIKEGIQKIREFAEQ----
Q8ZW57          RIALVEPEEQLAEAFSVLNEV-------
Q982E0          RLSYGIPTDQLVSGLARLTRLLNSWN--
Q982E3          RLSFGVPQDLLELGLKRVVETLNSLK--
O66630          RISLTVPTERLLEAAERIKNL-------
AAT2_METJA      RISYANSYENIKEGLERIKEFLNK----
O29838          RFALTRGEGVIEEAIDRLKTILHK----
Q9HQK2          RASFAVSTDAVREAVARIDAVLASATTL
AAT1_BACSU      RLSYATSLDLLEEAIERIKRFVEKH---
Q9YE99          RLSFATETERIREGIGRLARLLE-----
Q9R6Q3          RLSYATSLETLEAAVTRLKDWMND----
Q97AE8          RFAFSADTEHVREASELIEK--------
O30304          RISYANSLENLKKAAEIIREYESQR---
Q92QJ6          RFSYAGAATEMEEAMNRIARWL------
```

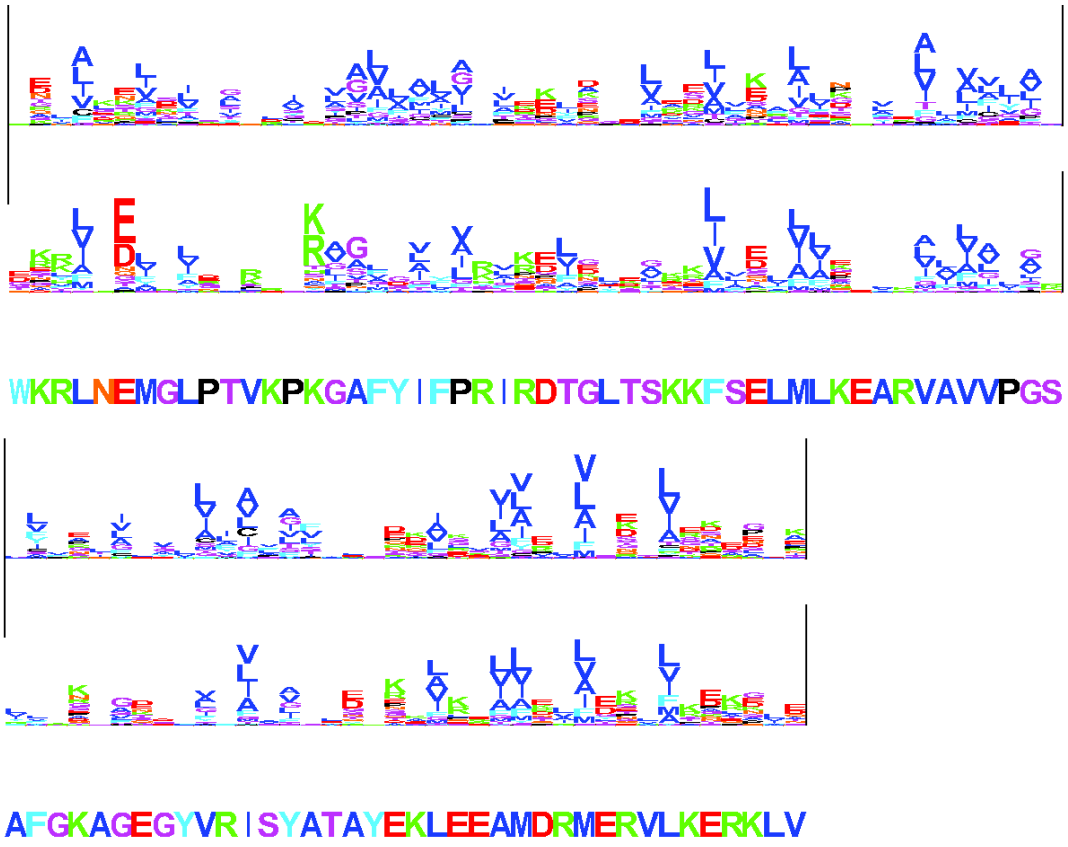
```
AAT_PYRHO       RISYATSKEKLIEAMNRIEKALEGEK--
Q98KW9          RISLCQPEPVLQEAAARLRRFASTYR--
Q8VS38          RISYATS----------------------
Q8XMH8          RACYAASMEDIMEAIKRIRRFVERNNM-
AAT_PYRAB       RISYATAYEKLEEAMDRMEKVLREKKL-
Q8VS39          ----------------------------
Q9KE01          RVALLADIAKLEEAMGRVGKL-------
O05237          ----------------------------
AAT_AQUAE       RLSYALSEERLVEGIRRIKKALEE----
Q8REF4          RFSFCVGENDIEKAMRVLEEALKVYK--
Q9PPF7          RLSYATSDELIKKGLERIANFIKNYK--
Q8Y1I0          RLSYATAMDHLEEAVARLARLFR-----
Q8RR70          RLSYATDLDTIKRGMERLEKFLH-----
Q9HRX4          RISYATDMATLRDAIDVMRAATAAVQ--
Q97M25          RISYAASMKDLKEGMIRLRRFIQSVKL-
Q9P9M8          RAVFLPPIEILEEAMDRFEKFMKER---
AAT_THEAQ       RLSYATGEENLKKALERFAQALQ-----
Q8XXV2          RFCFAKKDETLRLALERLSRL-------
Q8TS80          RCAYAASLDDIRKSIERMGDFVEELK--
Q8VXZ8          RFAFCKDEETLRGAIERMKQKLKRK---
Q8XRB5          RMSIATALETLEEGCRRIERAVLA----
```

B.5 Fichier additionnel 5

Marginal and leave-one-out profiles of complete protein partially displayed in figure 5

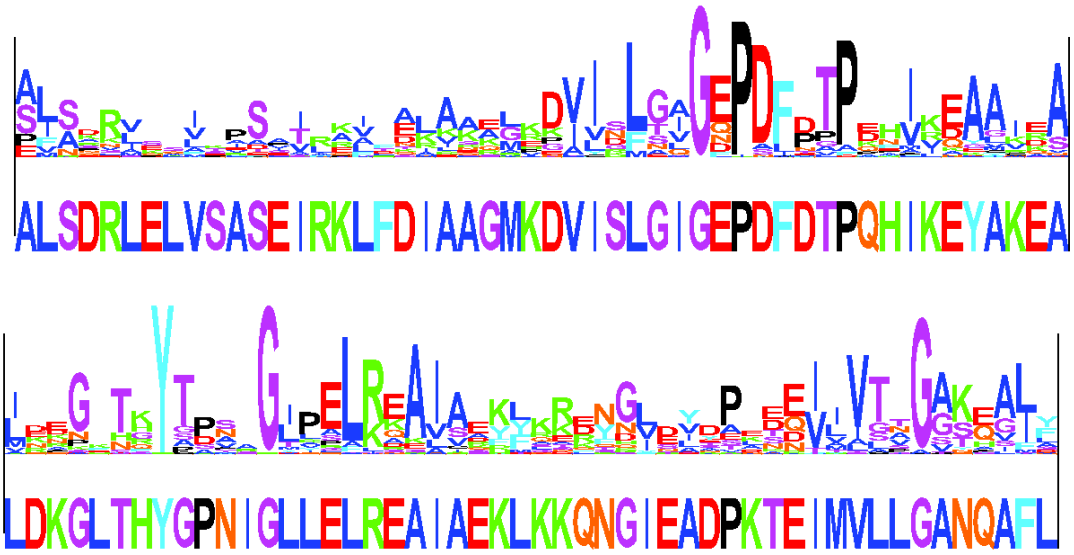


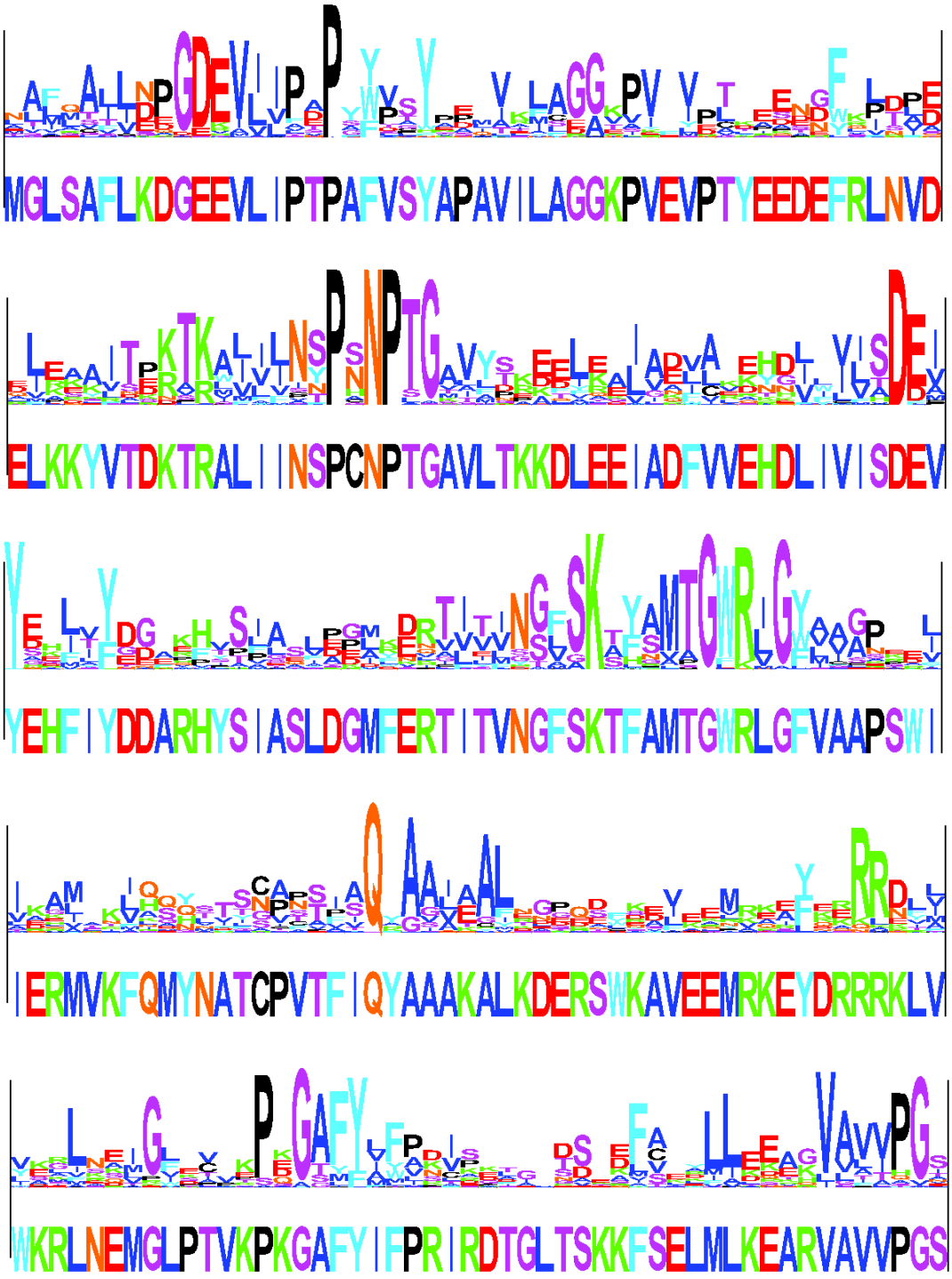




B.6 Fichier additionnel 6

Empirical profiles of complete protein partially displayed in figure 5

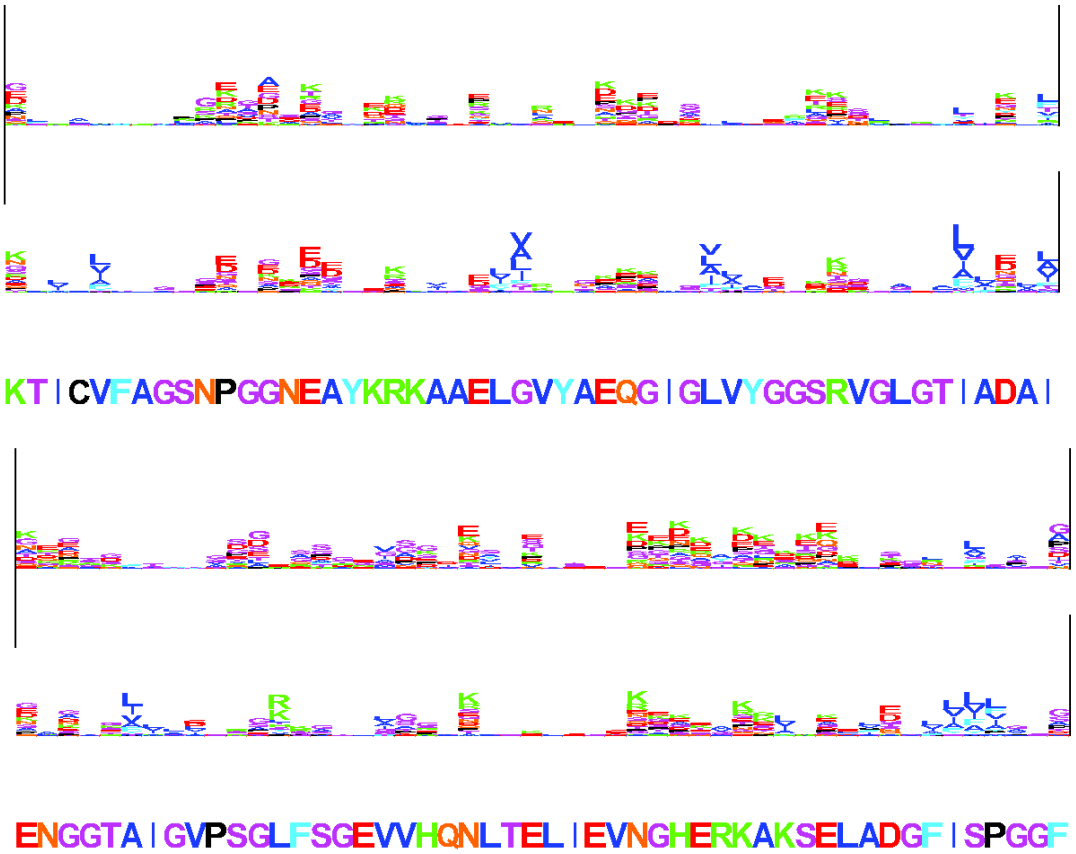


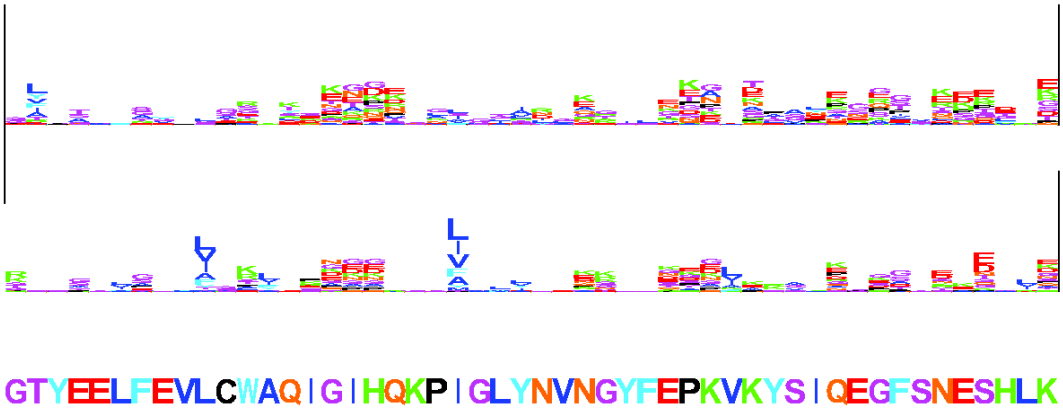

MGLSAFLKDGEEVLIPTPAFVSYAPAVILAGGKPVEVPTYEEDEFRLNVD
ELKKYVTDKTRALIINSPCNPTGAVLTKKDLEEIADFVVEHDLIVISDEV
YEHFIYDDARHYSIASLDGMFERTITVNGFSKTFAMTGWRLGFVAAPSWI
IERMVKFQMYNATCPVTFIQYAAAKALKDERSWKAVEEMRKEYDRRRKLV
WKRLNEMGLPTVKPKGAFYIFPRIRDTGLTSKKFSELMLKEARVAVVPGS

B.7 Fichier additionnel 7

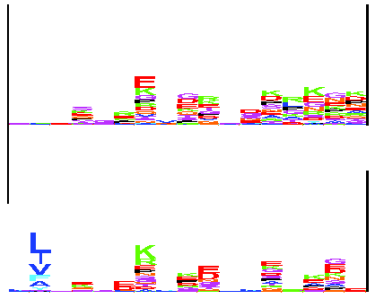
Marginal and leave-one-out profiles of 10 proteins used in the design specificity experiment

1T35A



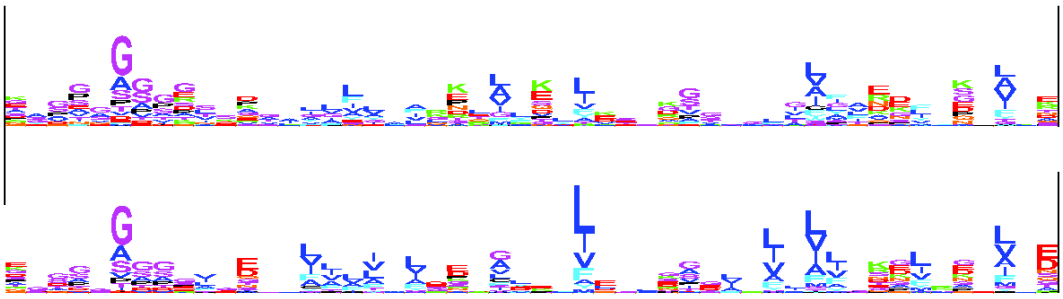


GTYEELFEVLCWAQ | G | HQKP | GLYNVNGYFEPKVKYS | QEGFSNESHLK

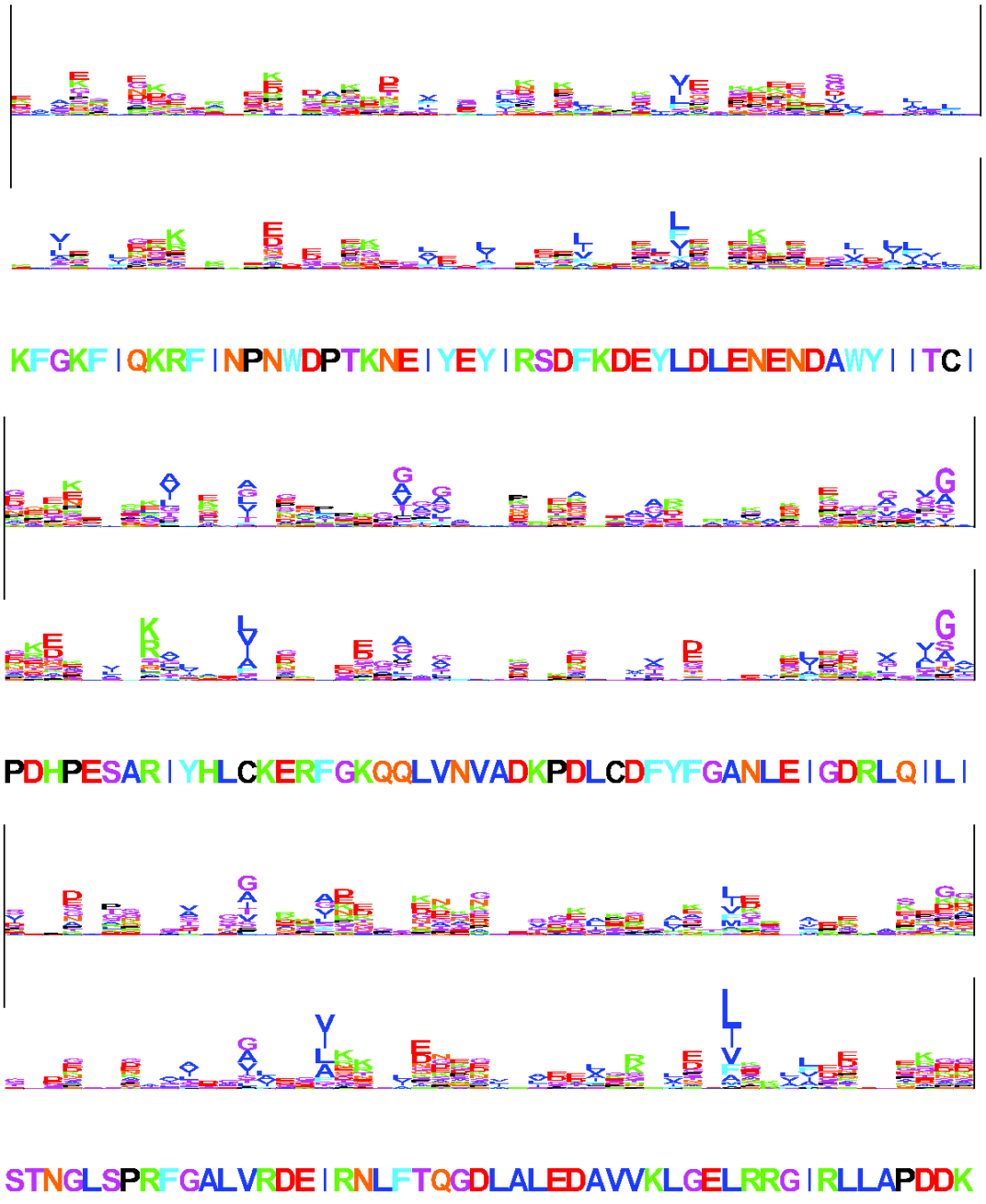


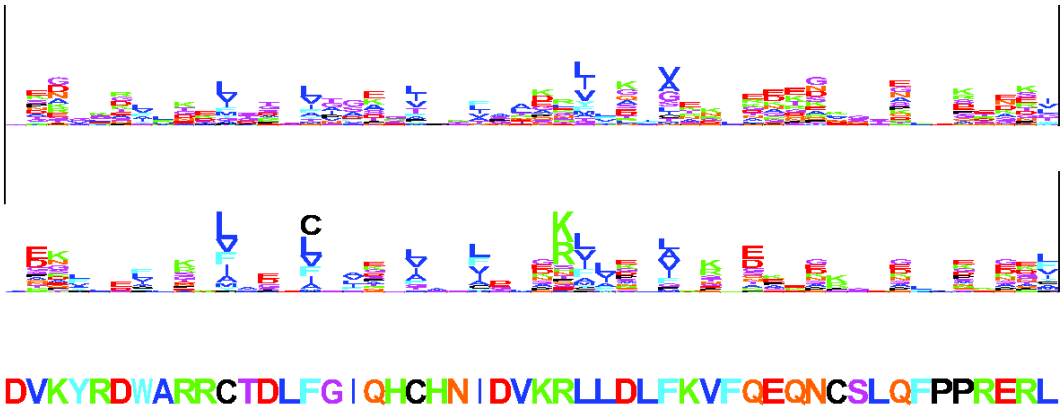
L | HSSSRPDEL | EQQNY

1KYQA

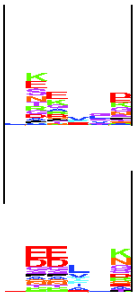


VKSLQLAHQLKDKR | LL | GGGEVGLTRLYKLPTGCKLTLVSPDLHKS | | P


KFGKF | QKRF | NPNWDPTKNE | YEY | RSDFKDEYLDLENENDAWY | | TC |
PDHPESAR | YHLCKERFGKQQLVNVADKPDLCDFYFGANLE | GDRLQ | L |
STNGLSPRFGALVRDE | RNLFTQGDLALEDAVVKLGELRRG | RLLAPDDK

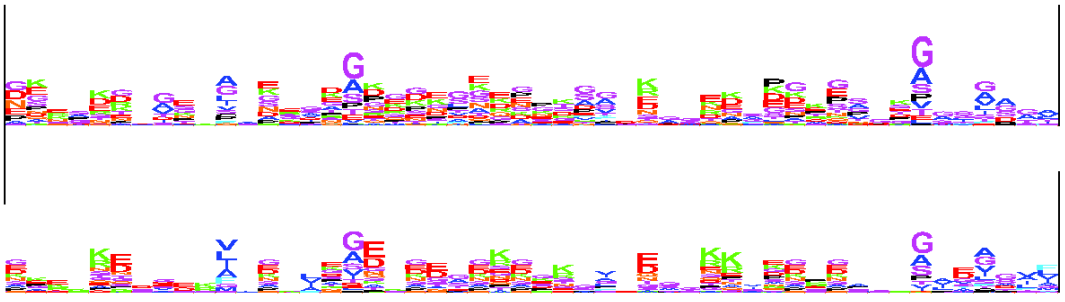


DVKYRDWARRCTDLFG | QHCHN | DVKRLLDLFKVFQEQNCSLQFPPRERL

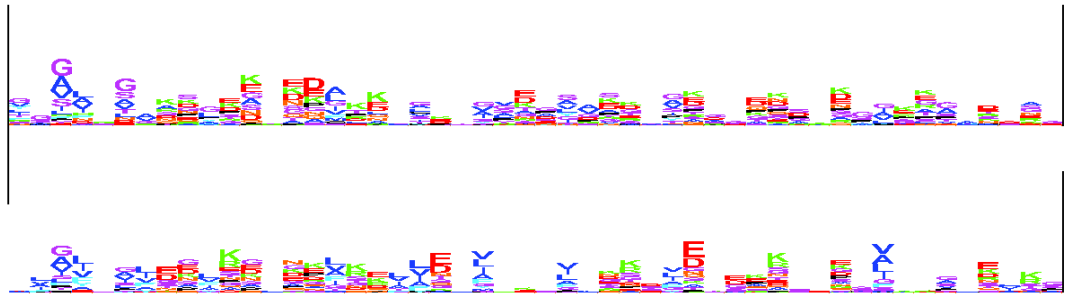


LSEYCS

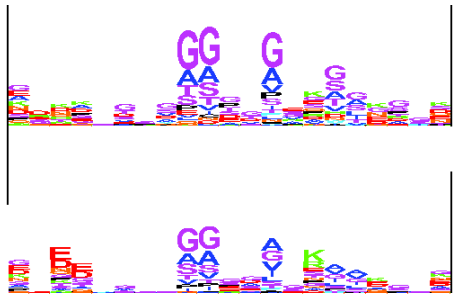
1D9CA



QGQFFRE | ENLKEYFNASSPDVAKGGPLFSE | LKNWKDESDKK | | QSQ | V

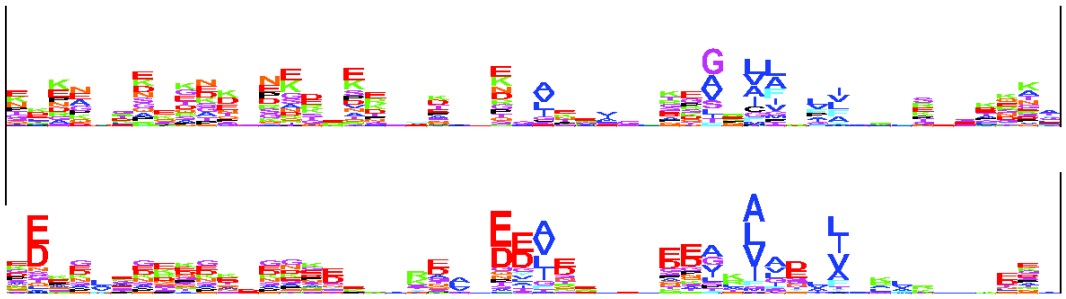


SFYFKLFENLKDNQV | QRSMD | | KQDMFQKFLNGSSEKLEDFKKL | Q | PV

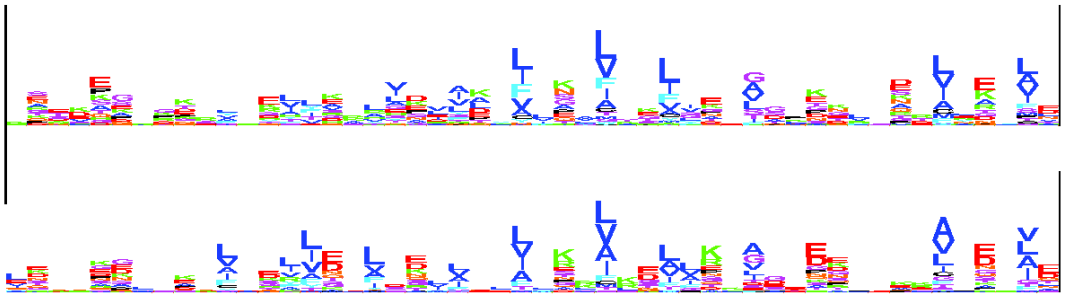


DDLQ | QRKA | NEL | KVMNDLS

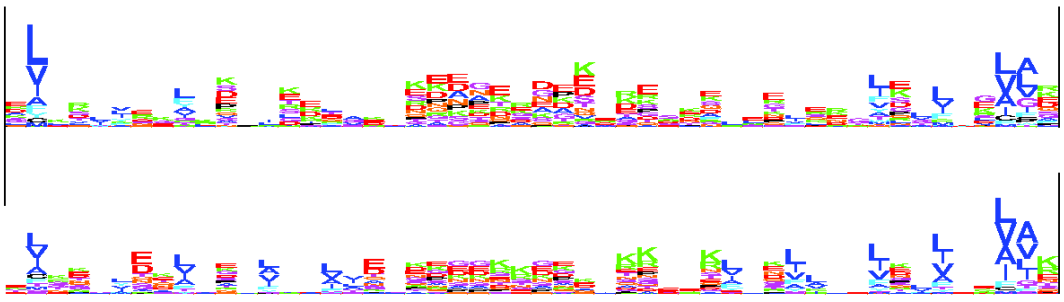
1QKRA



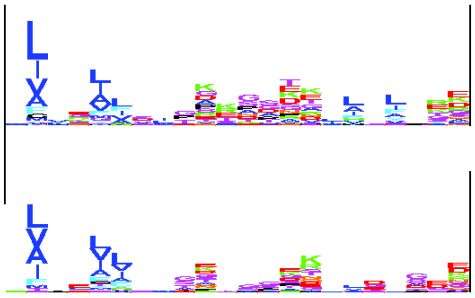
KDEEFPEQKAGEA | NQPAARQLHDEARKWSSKGND | | AAAKRALLAESRL



VRGGSGNKRAL | QCAKD | AKASDEVTRLAKEVAKQCTDKR | RTNLLQVCE



R | PT | STQLK | LSTVKATLGRTN | SDEESEQATELVHNAQNLQSVKETVR

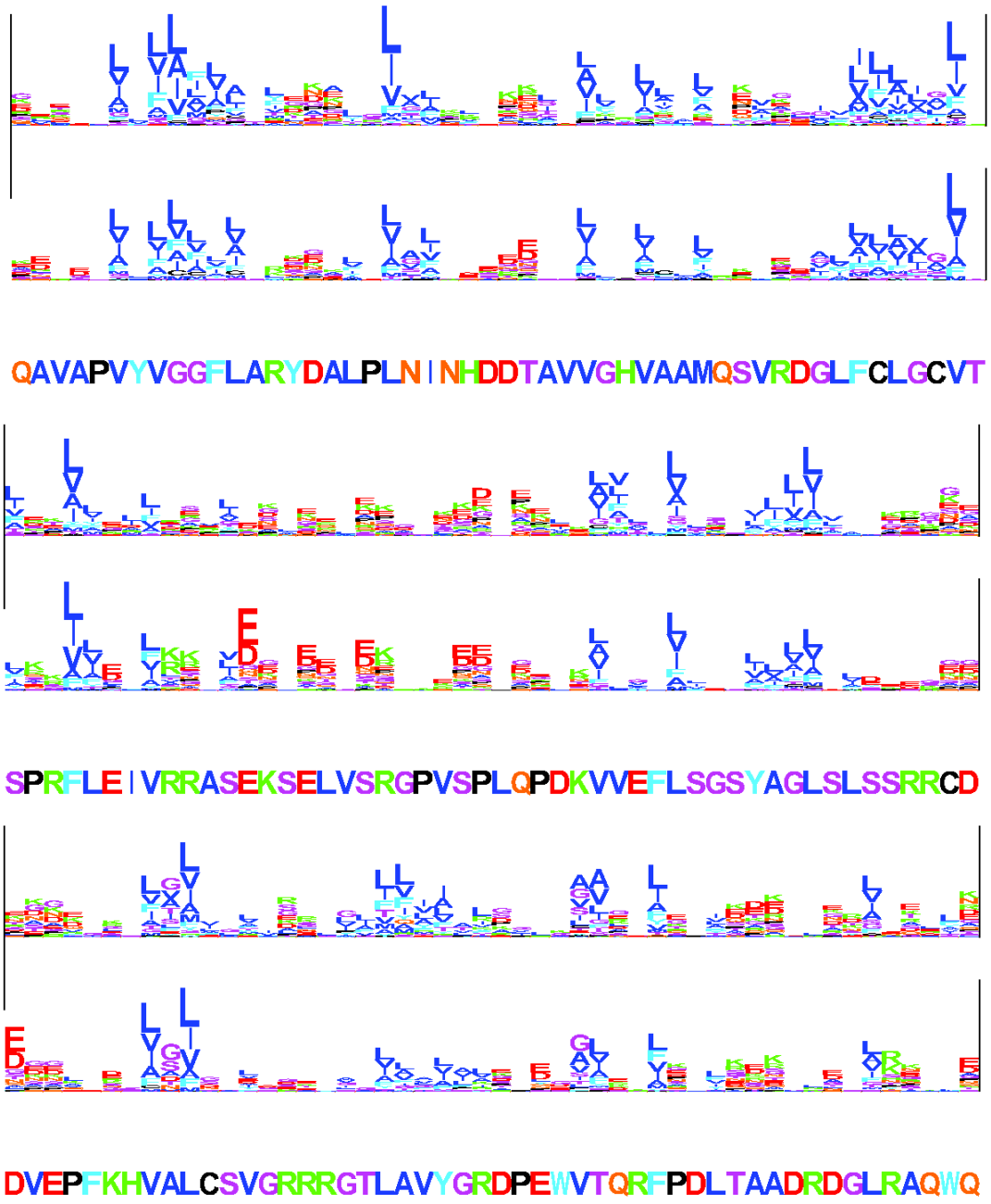


EAEAAS | K | RTDAGFTLRWVRK

1SFXA


HSNPLGELVKALEKLSFKPSDVR | YSLLLERGGRVSE | ARELDLSARFVR
DRLKVLLKRGFVRRE | VEKGWVGY | YSAEKPEKVLKEFKSS | LGE | ER | E
KFTDGS

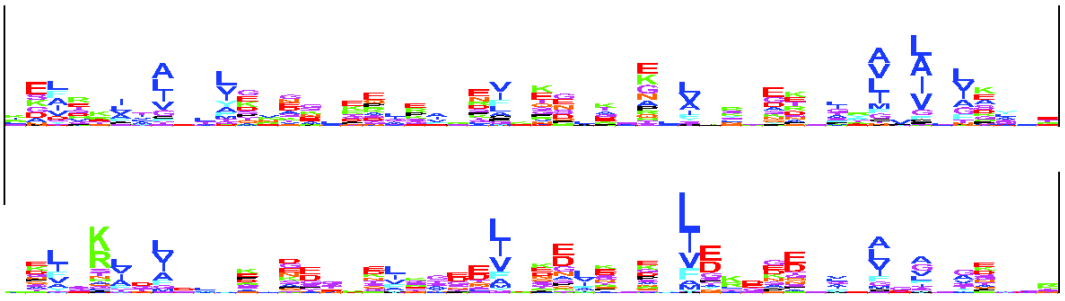
1LAY0


QAVAPVYVGGFLARYDALPLNINHDDTAVVGHVAAMQSVRDGLFCLGCVT
SPRFLEIVRRASEKSELVSRGPVSPLQPDKVVEFLSGSYAGLSLSSRRCD
DVEPFKHVALCSVGRRRGTLAVYGRDPEWVTQRFPDLTAADRDGLRAQWQ

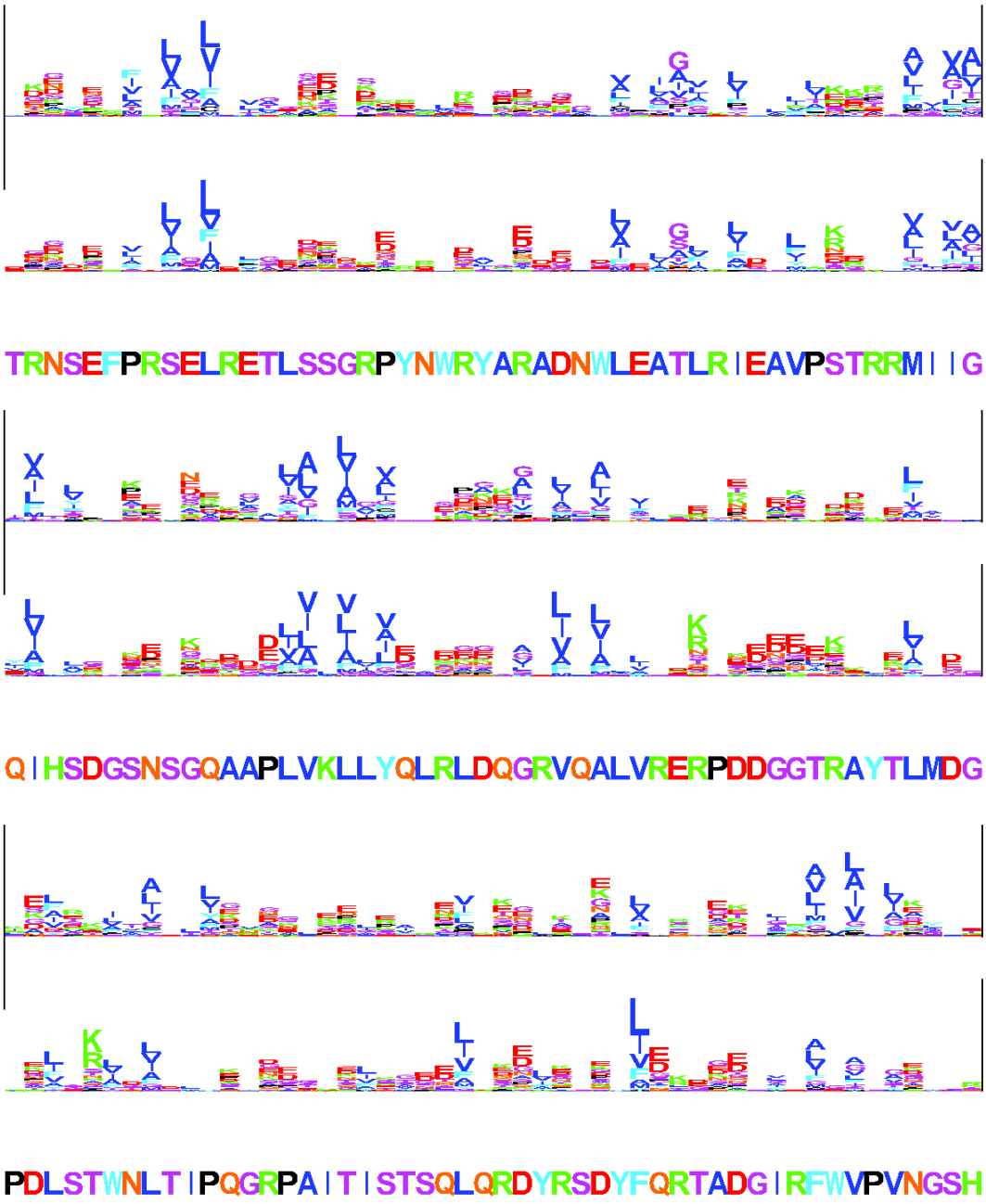
RCGSASGDPFRSDSYGLLGNSVDALY | RERLPKLRYDKQLVGVTERESYV

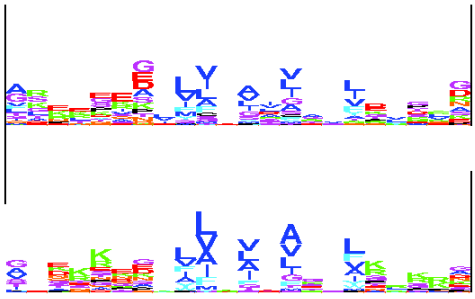
KA

1VAVA



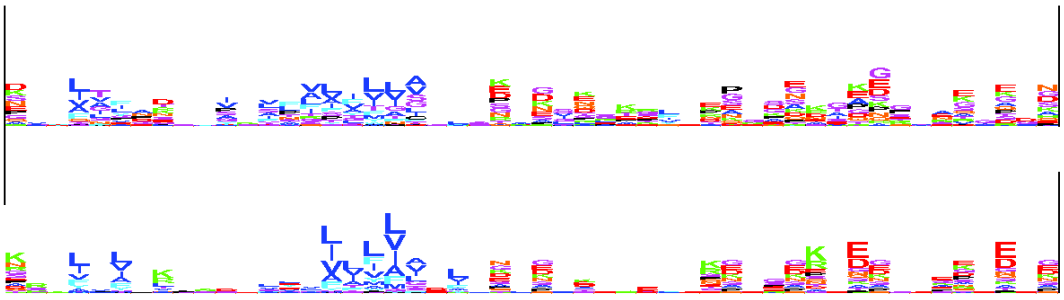
PDLSTWNLT | PQGRPA | T | STSQLQRDYRSDYFQRTADG | RFWVPVNGSH


TRNSEFPRSELRETLSSGRPYNWRYARADNWLEATLRIEAVPSTRRMIIG
QIHSDGSNSGQAAPLVKLLYQLRLDQGRVQALVRERPDDGGTRAYTLMDG
PDLSTWNLTIPQGRPAITISTSQLQRDYRSDYFQRTADGIRFWVPVNGSH

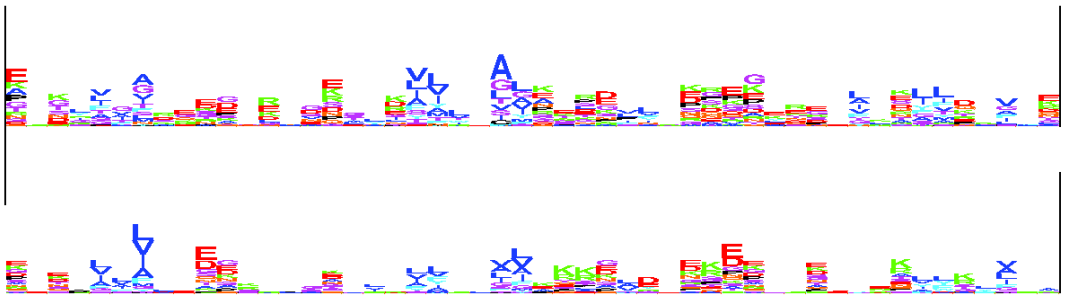


DNRGPSSEGGRATFSELRVSHQ

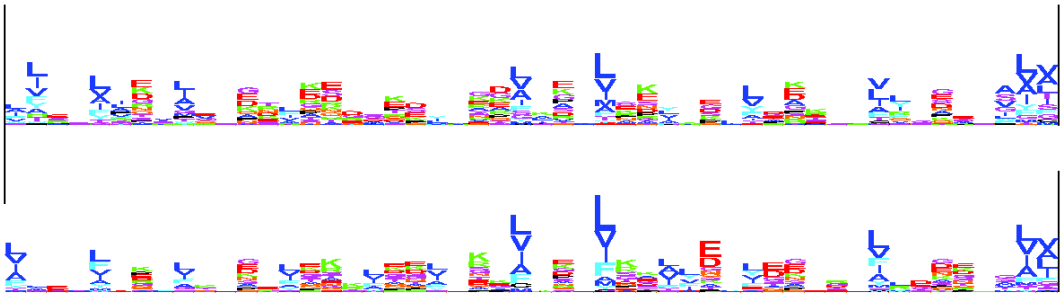
1B8XA



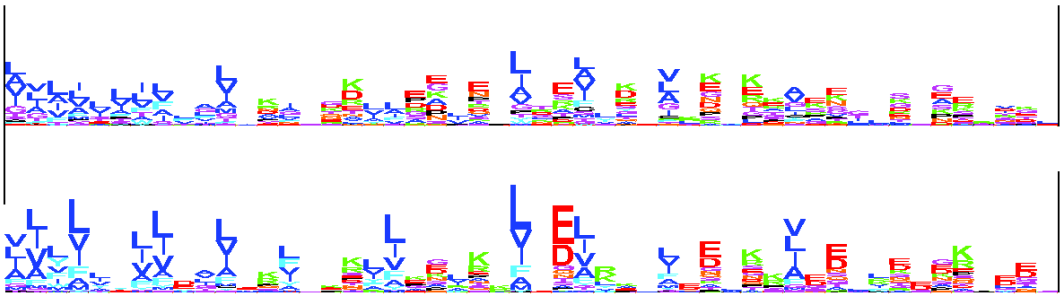
SP | LGYWK | KGLVQPTRLLLEYLEEKYEEHLYERDEGDKWRNKKFELGLE



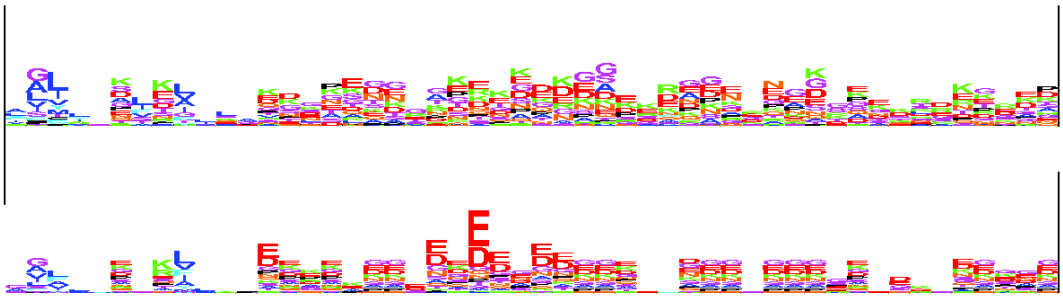
FPNLPYY | DGDVKLTQSMA | | RY | ADKHNMLGGCPKERAE | SMLEGAVLD



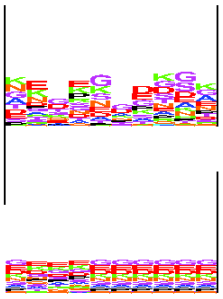
IRYGVSRIAYSKDFETLKVDFLSKLPEMLKMFEDRLCHKTYLNGDHVTHP



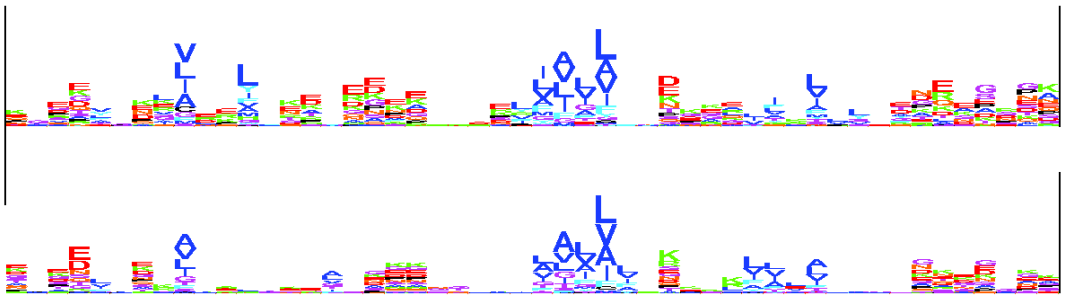
DFMLYDALDVVLYMDPMCLDAFPKLVCFKKRIEAIPQIDKYLKSSKYIAW

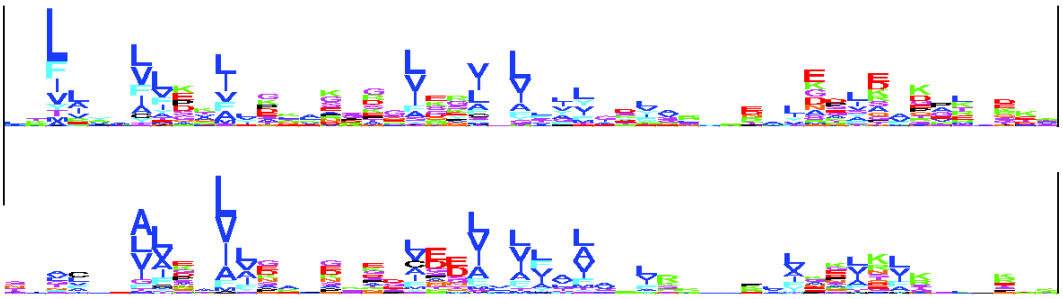


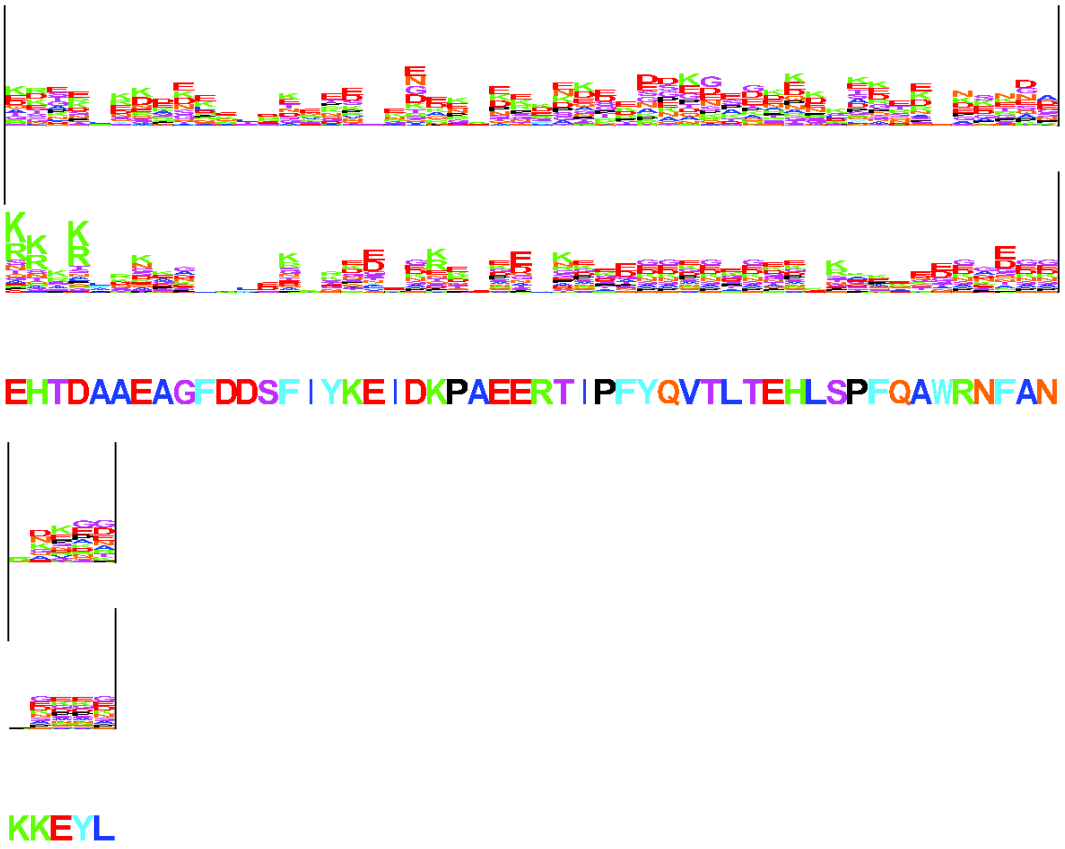
PLQGWQATFGGGDHPPKSDLVPRGSRRASVGSRMHYPGAFTYSPTPVTSG



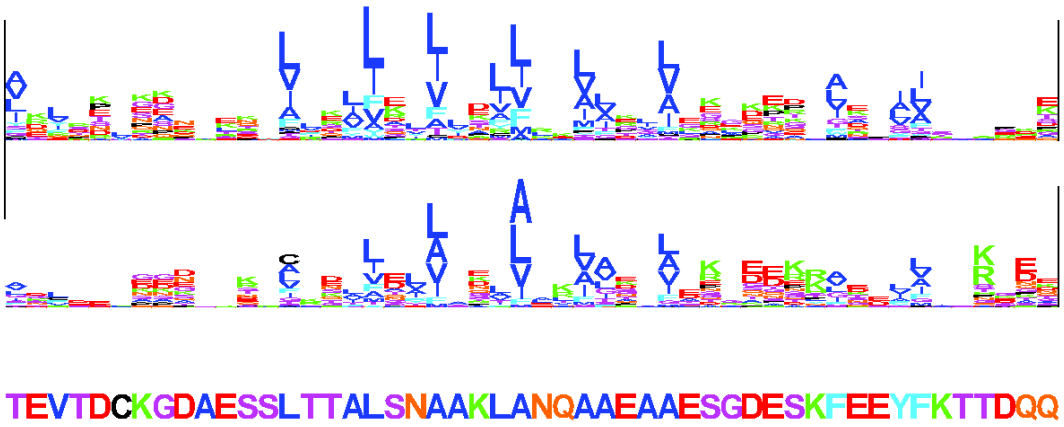
1TIYA

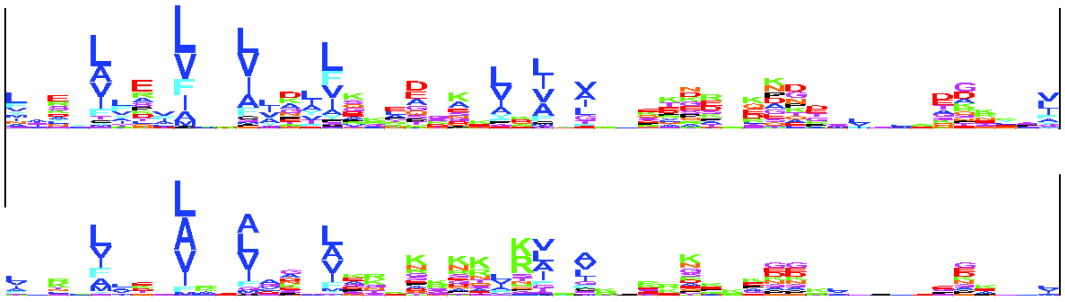




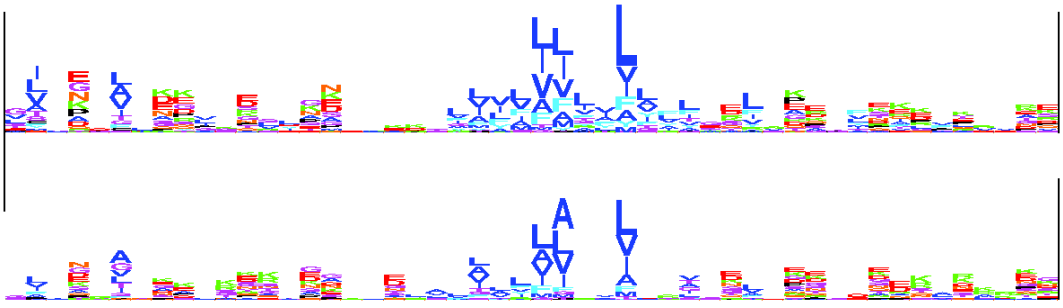


1EB6A

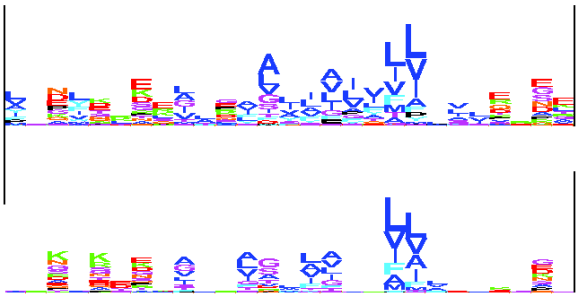




TRTTVAERLRAVAKEAGSTSGGSTTYHCNDPYGYCEPNVLAYTLPSKNE I



ANCD I YYSELPPLAQKCHAQDQATTTLHEFTHAPGVYQPGTEDLGYGYDA



ATQLSAQDALNNADSYALYANA I ELKC

B.8 Fichier additionnel 8

Table 1 : list of PDB identifiers of proteins used in the design specificity experiment, and scores obtained for each one of the proteins, using the combined $(\varepsilon + \alpha_{14ac} + \mu)$ ***potential***

PDB id	Average Z-score ratio	Ranking (median)	Target fold in top 1 %	Target fold in top 10 %	Average entropy/site	Average seq. Similarity	Average seq. Identity
1EB6A	0.8531	1	90 %	95 %	0.842	20.0	6.86
1TIYA	0.7978	12.5	75 %	100 %	0.766	17.4	6.61
1B8XA	0.7588	6.5	70 %	100 %	0.745	20.8	7.35
1VAVA	0.7143	1	90 %	100 %	0.742	22.9	7.91
1LAY0	0.7503	2.5	90 %	100 %	0.805	22.2	8.69
2SCUA	0.6885	1	90 %	100 %	0.838	22.5	8.37
1D2TA	0.7174	2	90 %	100 %	0.932	21.0	7.66
1Q4MA	0.5926	2.5	75 %	90 %	0.693	22.0	7.18
1B5L0	0.7181	2	80 %	100 %	0.766	25.4	8.85
1UMHA	0.5743	50	50 %	75 %	0.712	21.0	7.58
1NNGA	0.6736	2.5	85 %	95 %	0.755	22.7	7.16
1CE7A	0.6346	11.5	75 %	90 %	0.769	21.4	7.95
2PTH0	0.6749	6	100 %	100 %	0.821	21.1	8.03
1NAL3	0.6030	2.5	90 %	100 %	0.845	24.4	8.49
1CTT0	0.6205	1	100 %	100 %	0.862	23.2	8.69
1GS5A	0.5770	1	95 %	100 %	0.799	25.0	9.40
1DQYA	0.6078	5.5	90 %	100 %	0.814	21.1	7.53
1C8OA	0.5422	8	65 %	85 %	0.738	24.1	7.78
1VI9A	0.5486	2.5	80 %	100 %	0.765	21.8	8.12
1CFZA	0.5641	8.5	60 %	95 %	0.742	26.6	9.51
1UOX0	0.5280	5	65 %	90 %	0.700	21.4	7.34
1JJFA	0.5664	6.5	80 %	100 %	0.782	21.6	8.25
1D2NA	0.5389	5	85 %	100 %	0.767	26.2	8.72
1Q77A	0.5027	6.5	65 %	85 %	0.691	25.5	8.94
3CLA0	0.5053	15	65 %	90 %	0.665	21.9	7.19
5NUL0	0.5286	2	75 %	100 %	0.787	26.6	8.22
1CV80	0.5249	14	60 %	100 %	0.737	22.2	7.04
1B9LA	0.5373	48.5	50 %	75 %	0.716	22.3	7.02
1GHEA	0.5285	71	45 %	80 %	0.650	21.4	7.98
1D4AA	0.4495	78	35 %	75 %	0.725	21.8	8.32
1O70A	0.4782	6	75 %	95 %	0.743	24.1	7.64
1MUN0	0.4379	40.5	50 %	75 %	0.762	21.6	7.47
1RXQA	0.4393	37	55 %	80 %	0.725	24.2	8.25
1RIFA	0.4916	6	95 %	95 %	0.725	21.1	6.68
1D2ZB	0.4398	66.5	30 %	85 %	0.777	23.3	7.67
1HUW0	0.4336	98	35 %	80 %	0.734	23.4	7.14
3SDHA	0.3600	113.5	40 %	75 %	0.709	25.6	8.69
1H31A	0.4460	335	10 %	55 %	0.698	20.3	7.12
1EFDN	0.3922	11	70 %	100 %	0.712	22.5	7.54
1JJVA	0.3638	28.5	55 %	70 %	0.681	23.7	7.76

PDB id	Average Z-score ratio	Ranking (median)	Target fold in top 1 %	Target fold in top 10 %	Average entropy/site	Average seq. Similarity	Average seq. Identity
1RU8A	0.3374	178.5	10 %	70 %	0.678	23.0	8.54
1B74A	0.2983	89	50 %	53 %	0.801	26.0	9.35
1PA7A	0.3455	147	21 %	58 %	0.704	22.3	7.28
1EF8A	0.3580	95.5	30 %	85 %	0.778	22.6	7.87
1PQ4A	0.3762	54	40 %	85 %	0.756	24.7	7.88
1ETB1	0.3255	274	30 %	55 %	0.702	21.0	8.69
1ETEA	0.4506	405	40 %	50 %	0.717	23.1	7.31
3DFR0	0.3466	129	32 %	74 %	0.662	21.8	7.41
1UIZA	0.3600	176	30 %	70 %	0.726	21.3	8.09
3TMKA	0.3458	121.5	29 %	79 %	0.683	20.6	6.69
1AUVA	0.3139	114.5	35 %	70 %	0.641	21.2	7.06
1GGGA	0.2697	299.5	25 %	60 %	0.693	22.2	7.89
1F45B	0.1965	736	16 %	32 %	0.687	21.6	6.64
2SAK0	0.1740	629.5	15 %	40 %	0.701	22.7	8.80
1R6FA	0.1509	346	15 %	50 %	0.733	27.1	8.68
1T35A	0.1739	1326	5 %	26 %	0.571	19.0	6.16
1KYQA	-0.330	7 1056	8 %	28 %	0.660	21.3	6.97
1D9CA	0.0005	1558	0 %	15 %	0.731	23.8	6.57
1QKRA	0.0127	1878	5 %	15 %	0.727	23.7	7.76
1SFXA	-0.0508	2246	0 %	5 %	0.722	25.5	8.77
TOTAL	0.4526	32.75 53.6 %	77.5 %	0.738	22.7	7.82	

Annexe C

Optimisation des potentiels à l'aide d'une pseudo-vraisemblance

C.1 Fichier additionnel 1

Derivatives of the potential parameters

We found the derivative of the gradient :

$$\frac{\partial\omega(\tilde{s}|\tilde{s},c,\theta)}{\partial\theta} = \sum_{i=1..n}\left(-\frac{\partial F(\tilde{s}_i|\tilde{s}_{\setminus i},c,\theta)}{\partial\theta} + \sum_{a=1..20} p_i(a)\frac{\partial F(a|\tilde{s}_{\setminus i},c,\theta)}{\partial\theta}\right). \tag{C.1}$$

As for the joint criterion, the derivatives can be immediately calculated and become :

$$\frac{\partial\omega(\tilde{s}|\tilde{s},c)}{\partial\varepsilon_{ab}} = -n_{ab} + \sum_{1\le i<j\le n}\Delta_{ij}\left(P_i(a)\cdot D(b,s_j) + P_i(b)\cdot D(a,s_j)\right), \tag{C.2}$$

$$\frac{\partial\omega(\tilde{s}|\tilde{s},c,\theta)}{\partial\alpha_a^d} = -l_a^d + \sum_{1\le i\le n} P_i(a)D'(d,\nu_i), \tag{C.3}$$

where n_{ab} is the number of contacts between amino acids a and b observed in the database and l_a^d is the number of amino acids of type a belonging to the accessibility class d. $D(k,l) = 1$ if k and l are the same amino acid, and $D'(d,e) = 1$ if d and e are the same accessibility class. We can see these equations as

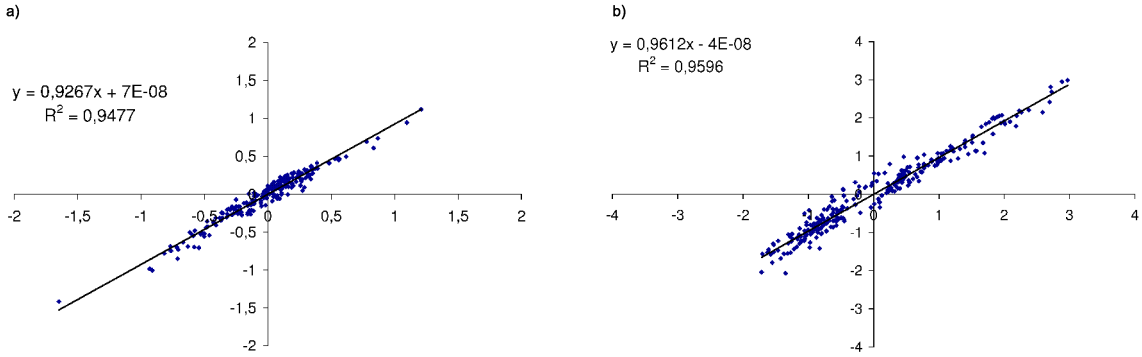
$$\frac{\partial\omega(\tilde{s}|\tilde{s},c,\theta)}{\partial\varepsilon_{ab}} = -n_{ab} + \langle n_{ab}\rangle_l, \tag{C.4}$$

$$\frac{\partial\omega(\tilde{s}|\tilde{s},c,\theta)}{\partial\alpha_a^d} = -l_a^d + \langle l_a^d\rangle_l, \tag{C.5}$$

which are the same formulation that were used for the derivatives for the joint criterion.

C.2 Fichier additionnel 2

XY-comparison of the leave-one-out potentials estimated from two independent datasets : (a) and (b) two independent runs on DS1 (X-axis) and DS2 (Y-axis) for contact and accessibility potentials respectivly.



C.3 Fichier additionnel 3

Contact potentials and solvent accessibility potentials written in an alphabetical order.

Contact potential parameters

amino acid order : A C D E F G H I K L M N P Q R S T V W Y

-0.0881964 0.172555 0.348397 0.336866 -0.0444176 0.336243 0.33816 -0.0774892 0.435863 -0.245584 -0.0746802 0.32986 0.363204 0.241991 0.244724 0.33735 0.196917 -0.0601244 0.108651 0.118756

0.172555 -1.27117 0.344291 0.351299 -0.246531 0.313573 -0.0412198 -0.102118 0.246814 -0.261753 -0.184324 0.212077 0.107969 0.148392 0.211543 0.26443 0.19936 -0.0440262 -0.0290509 -0.00121993

0.348397 0.344291 0.410586 0.631294 0.297549 0.320746 -0.314693 0.329846 -0.688834 0.221213 0.203352 -0.114669 0.460178 0.00292988 -0.778366 0.0680921 0.0637852 0.402581 0.231306 -0.101285

0.336866 0.351299 0.631294 0.472418 -0.0247011 0.518242 -0.370772 0.0136588 -0.888781 -0.151384 -0.0912826 -0.00820533 0.308829 -0.0559046 -1.00903 0.115381 0.0220514 0.145493 -0.0492177 -0.267574

-0.0444176 -0.246531 0.297549 -0.0247011 -0.781081 0.341467 -0.172872 -0.504444 -0.176614 -0.735123 -0.710012 0.0741938 -0.119722 -0.202443 -0.194501 0.209673 0.0478155 -0.371082 -0.414962 -0.544263

0.336243 0.313573 0.320746 0.518242 0.341467 0.234123 0.377636 0.437014 0.329582 0.351052 0.256122 0.191943 0.445118 0.36557 0.263024 0.349878 0.306363 0.394181 0.403634 0.344884

0.33816 -0.0412198 -0.314693 -0.370772 -0.172872 0.377636 -0.407206 0.0801043 0.365998 -0.173685 -0.278433 0.102949 0.242793 -0.0277533 0.104083 0.138677 0.0565216 0.150452 -0.185474 -0.316087

-0.0774892 -0.102118 0.329846 0.0136588 -0.504444 0.437014 0.0801043 -0.626812 -0.101027 -0.730964 -0.525032 0.159217 0.132889 -0.0747174 0.010633 0.3012 0.0137866 -0.432557 -0.130716 -0.319005

0.435863 0.246814 -0.688834 -0.888781 -0.176614 0.329582 0.365998 -0.101027 0.873698 -0.23944 -0.197305 -0.00478472 0.574953 -0.049736 0.763901 0.267216 0.151207 0.0645472 -0.166429 -0.401886

-0.245584 -0.261753 0.221213 -0.151384 -0.735123 0.351052 -0.173685 -0.730964 -0.23944 -1.00452 -0.693246 0.0690928 -0.0684994 -0.336799 -0.316371 0.190174 -0.112533 -0.596402 -0.398603 -0.556426

-0.0746802 -0.184324 0.203352 -0.0912826 -0.710012 0.256122 -0.278433 -0.525032 -0.197305 -0.693246 -0.973939 -0.0479707 -0.111912 -0.300612 -0.230563 0.198542 -0.0688106 -0.380274 -0.488581 -0.566624
0.32986 0.212077 -0.114669 -0.00820533 0.0741938 0.191943 0.102949 0.159217 -0.00478472 0.0690928 -0.0479707 -0.316095 0.288451 -0.115945 0.0783777 0.110392 -0.00642609 0.265101 0.0715755 -0.0711156
0.363204 0.107969 0.460178 0.308829 -0.119722 0.445118 0.242793 0.132889 0.574953 -0.0684994 -0.111912 0.288451 0.28344 0.156126 0.206253 0.425475 0.294396 0.177678 -0.266024 -0.249608
0.241991 0.148392 0.00292988 -0.0559046 -0.202443 0.36557 -0.0277533 -0.0747174 -0.049736 -0.336799 -0.300612 -0.115945 0.156126 -0.258259 -0.14435 0.144157 -0.0580422 0.0432564 -0.26747 -0.263551
0.244724 0.211543 -0.778366 -1.00903 -0.194501 0.263024 0.104083 0.010633 0.763901 -0.316371 -0.230563 0.0783777 0.206253 -0.14435 0.305648 0.277484 0.151483 0.0881696 -0.315524 -0.300311
0.33735 0.26443 0.0680921 0.115381 0.209673 0.349878 0.138677 0.3012 0.267216 0.190174 0.198542 0.110392 0.425475 0.144157 0.277484 0.205269 0.204309 0.333244 0.288517 0.245381
0.196917 0.19936 0.0637852 0.0220514 0.0478155 0.306363 0.0565216 0.0137866 0.151207 -0.112533 -0.0688106 -0.00642609 0.294396 -0.0580422 0.151483 0.204309 0.0153926 0.0549045 0.26597 0.134405
-0.0601244 -0.0440262 0.402581 0.145493 -0.371082 0.394181 0.150452 -0.432557 0.0645472 -0.596402 -0.380274 0.265101 0.177678 0.0432564 0.0881696 0.333244 0.0549045 -0.378807 -0.0333088 -0.173367
0.108651 -0.0290509 0.231306 -0.0492177 -0.414962 0.403634 -0.185474 -0.130716 -0.166429 -0.398603 -0.488581 0.0715755 -0.266024 -0.26747 -0.315524 0.288517 0.26597 -0.0333088 -0.319452 -0.243701
0.118756 -0.00121993 -0.101285 -0.267574 -0.544263 0.344884 -0.316087 -0.319005 -0.401886 -0.556426 -0.566624 -0.0711156 -0.249608 -0.263551 -0.300311 0.245381 0.134405 -0.173367 -0.243701 -0.350486

Solvent accessibility potential parameters : 14 classes (columns)
amino acid order : A C D E F G H I K L M N P Q R S T V W Y
-0.0881964 0.172555 0.348397 0.336866 -0.0444176 0.336243 0.33816 -0.0774892 0.435863 -0.245584 -0.0746802 0.32986 0.363204 0.241991 0.244724 0.33735 0.196917 -0.0601244 0.108651 0.118756
0.172555 -1.27117 0.344291 0.351299 -0.246531 0.313573 -0.0412198 -0.102118 0.246814 -0.261753 -0.184324 0.212077 0.107969 0.148392 0.211543 0.26443 0.19936 -0.0440262 -0.0290509 -0.00121993
0.348397 0.344291 0.410586 0.631294 0.297549 0.320746 -0.314693 0.329846 -0.688834 0.221213 0.203352 -0.114669 0.460178 0.00292988 -0.778366 0.0680921 0.0637852 0.402581 0.231306 -0.101285
0.336866 0.351299 0.631294 0.472418 -0.0247011 0.518242 -0.370772 0.0136588 -0.888781 -0.151384 -0.0912826 -0.00820533 0.308829 -0.0559046 -1.00903 0.115381 0.0220514 0.145493 -0.0492177 -0.267574
-0.0444176 -0.246531 0.297549 -0.0247011 -0.781081 0.341467 -0.172872 -0.504444 -0.176614 -0.735123 -0.710012 0.0741938 -0.119722 -0.202443 -0.194501 0.209673 0.0478155 -0.371082 -0.414962 -0.544263
0.336243 0.313573 0.320746 0.518242 0.341467 0.234123 0.377636 0.437014 0.329582 0.351052 0.256122 0.191943 0.445118 0.36557 0.263024 0.349878 0.306363 0.394181 0.403634 0.344884
0.33816 -0.0412198 -0.314693 -0.370772 -0.172872 0.377636 -0.407206 0.0801043 0.365998 -0.173685 -0.278433 0.102949 0.242793 -0.0277533 0.104083 0.138677 0.0565216 0.150452 -0.185474 -0.316087
-0.0774892 -0.102118 0.329846 0.0136588 -0.504444 0.437014 0.0801043 -0.626812 -0.101027 -0.730964 -0.525032 0.159217 0.132889 -0.0747174 0.010633 0.3012 0.0137866 -0.432557 -0.130716 -0.319005
0.435863 0.246814 -0.688834 -0.888781 -0.176614 0.329582 0.365998 -0.101027 0.873698 -0.23944 -0.197305 -0.00478472 0.574953 -0.049736 0.763901 0.267216 0.151207 0.0645472 -0.166429 -0.401886
-0.245584 -0.261753 0.221213 -0.151384 -0.735123 0.351052 -0.173685 -0.730964 -0.23944 -1.00452 -0.693246 0.0690928 -0.0684994 -0.336799 -0.316371 0.190174 -0.112533 -0.596402 -0.398603 -0.556426
-0.0746802 -0.184324 0.203352 -0.0912826 -0.710012 0.256122 -0.278433 -0.525032 -0.197305 -0.693246 -0.973939 -0.0479707 -0.111912 -0.300612 -0.230563 0.198542 -0.0688106 -0.380274 -0.488581 -0.566624
0.32986 0.212077 -0.114669 -0.00820533 0.0741938 0.191943 0.102949 0.159217 -0.00478472 0.0690928 -0.0479707 -0.316095 0.288451 -0.115945 0.0783777 0.110392 -0.00642609 0.265101 0.0715755 -0.0711156
0.363204 0.107969 0.460178 0.308829 -0.119722 0.445118 0.242793 0.132889 0.574953 -0.0684994 -0.111912 0.288451 0.28344 0.156126 0.206253 0.425475 0.294396 0.177678 -0.266024 -0.249608
0.241991 0.148392 0.00292988 -0.0559046 -0.202443 0.36557 -0.0277533 -0.0747174 -0.049736 -0.336799 -0.300612 -0.115945 0.156126 -0.258259 -0.14435 0.144157 -0.0580422 0.0432564 -0.26747 -0.263551
0.244724 0.211543 -0.778366 -1.00903 -0.194501 0.263024 0.104083 0.010633 0.763901 -0.316371 -0.230563 0.0783777 0.206253 -0.14435 0.305648 0.277484 0.151483 0.0881696 -0.315524 -0.300311

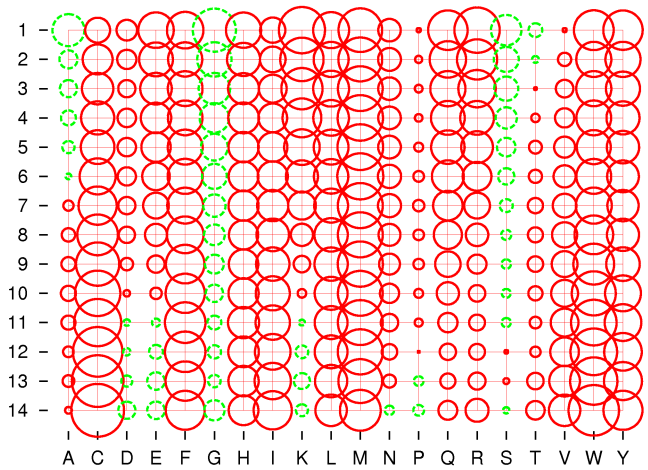
0.33735 0.26443 0.0680921 0.115381 0.209673 0.349878 0.138677 0.3012 0.267216 0.190174 0.198542 0.110392 0.425475 0.144157 0.277484 0.205269 0.204309 0.333244 0.288517 0.245381
0.196917 0.19936 0.0637852 0.0220514 0.0478155 0.306363 0.0565216 0.0137866 0.151207 -0.112533 -0.0688106 -0.00642609 0.294396 -0.0580422 0.151483 0.204309 0.0153926 0.0549045 0.26597 0.134405
-0.0601244 -0.0440262 0.402581 0.145493 -0.371082 0.394181 0.150452 -0.432557 0.0645472 -0.596402 -0.380274 0.265101 0.177678 0.0432564 0.0881696 0.333244 0.0549045 -0.378807 -0.0333088 -0.173367
0.108651 -0.0290509 0.231306 -0.0492177 -0.414962 0.403634 -0.185474 -0.130716 -0.166429 -0.398603 -0.488581 0.0715755 -0.266024 -0.26747 -0.315524 0.288517 0.26597 -0.0333088 -0.319452 -0.243701

0.118756 -0.00121993 -0.101285 -0.267574 -0.544263 0.344884 -0.316087 -0.319005 -0.401886 -0.556426 -0.566624 -0.0711156

-0.249608 -0.263551 -0.300311 0.245381 0.134405 -0.173367 -0.243701 -0.350486

C.4 Fichier additionnel 4

Bubble plot of the solvant accessibility potential where we remove from each potential the corresponding natural logarithm frequency of the accessibility class.



C.5 Fichier additionnel 5

Controlled inertial gradient algorithm

Algorithm 1: Choice algorithm for the gradient descent

Data: A set of values of parameters $\theta^{(m)}$, a gradient $d\Omega = \frac{\partial \omega(\tilde{s}|\tilde{s},c,\theta)}{\partial \theta}$, a step δ_{grad}, a value $0 < r < 1$.

Results: A set of values of parameters $\theta^{(m+1)}$ like $\omega^l(\tilde{s}|\tilde{s},c,\theta^{(m+1)}) < \omega^l(\tilde{s}|\tilde{s},c,\theta^{(m)})$.

begin

- $\theta^{(m+1)} := \theta^{(m)}$
- **while** $\theta^{(m+1)} = \theta^{(m)}$ **do**
  - $\theta^\star = \theta^{(m)} - \delta_{grad} d\Omega - \Delta\theta^{(m)}$
  - **if** $\omega^l(\tilde{s}|\tilde{s},c,\theta^\star) < \omega^l(\tilde{s}|\tilde{s},c,\theta^{(m)})$ **then**
    - $\theta^{(m+1)} = \theta^\star$
  - **else**
    - $\theta^\star = \theta^{(m)} + \delta_{grad} d\Omega$
    - **if** $\omega^l(\tilde{s}|\tilde{s},c,\theta^\star) < \omega^l(\tilde{s}|\tilde{s},c,\theta^{(m)})$ **then**
      - $\theta^{(m+1)} = \theta^\star$
    - **else**
      - $\delta_{grad} = \delta_{grad} * r$

end

C.6 Fichier additionnel 6

Inclusion of μ_a in the accessibility terms

Keeping the constraints defined by

$$\sum_{1 \leq a \leq 20} \mu_a = 0, \tag{C.6}$$

$$\sum_{1 \leq a \leq 20} \sum_{1 \leq b \leq 20} \varepsilon_{ab} = 0, \tag{C.7}$$

$$\sum_{1 \leq a \leq 20} \alpha_a^d = 0, d = \{1..D\}, \tag{C.8}$$

if we define

$$\forall d \quad \alpha'^d_{a_k} = \alpha^d_{a_k} + J \quad \text{and} \quad \mu'_{a_k} = \mu_{a_k} - J, \tag{C.9}$$

and

$$\forall d \quad \alpha'^{d}_{a_l} = \alpha^{d}_{a_l} - J \quad \text{and} \quad \mu'_{a_l} = \mu_{a_l} + J, \tag{C.10}$$

then

$$\begin{aligned} G'(s,c) &= \sum_{1\leq i<j\leq n} \varepsilon_{s_i s_j} + \sum_{1\leq i\leq n} \alpha'^{\nu_i}_{s_i} + \sum_{1\leq i\leq n} \mu'_{s_i} && \text{(C.11)} \\ &= \sum_{1\leq i<j\leq n} \varepsilon_{s_i s_j} + \sum_{1\leq i\leq n} \alpha^{\nu_i}_{s_i} - J + J + \sum_{1\leq i\leq n} \mu_{s_i} + J - J && \text{(C.12)} \\ &= G(s,c). && \text{(C.13)} \end{aligned}$$

So, repeting this iteratively from $(k = a_1, l = a_2)$ to $(k = a_{19}, l = a_{20})$, then

$$\forall a, 1 \leq a \leq 20, \quad \mu_a = 0. \tag{C.14}$$

And thus, $\mu_a, a = \{1..20\}$ terms can be all included in the accessibility terms.

Bibliographie

[Abkevich et al., 1996] Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1996). Improved design of stable and fast-folding model proteins. *Fold. Des.*, 1 :221–230.

[Adachi and Hasegawa, 1996] Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, 42 :459–468.

[Adachi et al., 2000] Adachi, J., Waddell, P. J., Martin, W., and Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.*, 50 :348–358.

[Aitkin, 1991] Aitkin, M. (1991). Posterior Bayes factors. *J. R. Stat. Soc. B*, 53(1) :111–142.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control*, AC-19(6) :716–723.

[Altekar et al., 2004] Altekar, G., Dwarkadas, S., Huelsenbeck, J., and Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3) :407–415.

[Anfinsen, 1973] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181 :223–230.

[Antoniak, 1974] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.*, 2 :1152–1174.

[Baldauf et al., 2000] Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, 290 :972–977.

[Banavar et al., 1998] Banavar, J., Cieplak, M., Maritan, A., Nadig, G., Seno, F., and Vishveshwara, S. (1998). Structure-based design of model proteins. *Proteins*, 32 :80–87.

[Bapteste et al., 2002] Bapteste, E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M., and Philippe, H. (2002). The analysis of 100 genes supports the grouping of three highly divergent amoebae : Dictyostelium, Entamoeba, and Mastigamoeba. *Proc. Natl. Acad. Sci. U.S.A.*, 99 :1414–1419.

[Bastolla et al., 2001] Bastolla, U., Farwer, J., Knapp, E. W., and Vendruscolo, M. (2001). How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins*, 44 :79–96.

[Bastolla et al., 2008a] Bastolla, U., Ortíz, A. R., Porto, M., and Teicher, F. (2008a). Effective connectivity profile : A structural representation that evidences the relationship between protein structures and sequences. *Proteins*, 73(4) :872–888.

[Bastolla et al., 2008b] Bastolla, U., Porto, M., and Ortíz, A. R. (2008b). Local interactions in protein folding determined through an inverse folding potential. *Proteins*, 71 :278–299.

[Bastolla et al., 2002] Bastolla, U., Porto, M., Roman, H. E., and Vendruscolo, M. (2002). Lack of self-averaging in neutral evolution of proteins. *Phys. Rev. Lett.*, 89.

[Bastolla et al., 2003] Bastolla, U., Porto, M., Roman, H. E., and Vendruscolo, M. (2003). Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *J. Mol. Evol.*, 56 :243–254.

[Bastolla et al., 2006] Bastolla, U., Porto, M., Roman, H. E., and Vendruscolo, M. (2006). A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the protein data bank. *BMC Evol. Biol.*, 6 :43.

[Bastolla et al., 1999] Bastolla, U., Roman, H. E., and Vendruscolo, M. (1999). Neutral evolution of model porteins : diffusion in sequence space and overdispersion. *J. Theor. Biol.*, 200 :49–64.

[Bastolla et al., 2000] Bastolla, U., Vendruscolo, M., and Knapp, E. W. (2000). A statistical mechanical method to optimize energy functions for protein folding. *Proc. Natl. Acad. Sci. USA*, 97(8) :3977–3981.

[Basu and Chib, 2003] Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factoes for Dirichlet process mixture models. *J. Am. Stat. Assoc.*, 98 :224–235.

[Bauer and Beyer, 1994] Bauer, A. and Beyer, A. (1994). An improved pair potential to recognize native protein folds. *Proteins*, 18 :254–261.

[Berger and Pericchi, 1996] Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Am. Stat. Assoc.*, 91 :109–122.

[Betancourt and Skolnik, 2004] Betancourt, M. R. and Skolnik, J. (2004). Local propensities and statiscal potentials of backbone dihedral angles in proteins. *J. Mol. Biol.*, 342 :635–649.

[Betancourt and Thirumalai, 1999] Betancourt, M. R. and Thirumalai, D. (1999). Pair potentials for protein folding : Choice of reference states and sensitivity of predicted mative states to variations in the interaction schemes. *Protein Sci.*, 8 :361–369.

[Blanquart, 2007] Blanquart, S. (2007). *Reconstruction phylogénétique par analyse Bayésienne des séquences moléculaires*. PhD thesis, Université de Montpellier 2.

[Bollback, 2002] Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, 19 :1171–1180.

[Bolon et al., 2005] Bolon, D. N., Grant, R. A., Baker, T. A., and Sauer, R. T. (2005). Specificity versus stability in computational protein design. *Proc. Natl. Acad. Sci. USA*, 36 :12724–12729.

[Bonnard et al., 2006] Bonnard, C., Berry, V., and Lartillot, N. (2006). Multipolar consensus for phylogenetic trees. *Syst. Biol.*, 55(5) :837–843.

[Bonnard et al., 2009] Bonnard, C., Kleinman, C. L., Rodrigue, N., and Lartillot, N. (2009). Fast optimization of statistical potentials for structurally constrained phylogenetic models. *BMC Evol. Biol.*, 9 :227.

[Bos, 2002] Bos, C. S. (2002). A comparison of marginal likelihood computation methods. Tinbergen institute discussion papers, Tinbergen Institute.

[Bourque et al., 2004] Bourque, G., Pevzner, P. A., and Tesler, G. (2004). Reconstructing the genomic architecture of ancestral mammals : lessons from human, mouse and rat genomes. *Genome Res.*, 14 :507–516.

[Bowie et al., 1991] Bowie, J. U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016) :164–170.

[Branden and Tooze, 1999] Branden, C. and Tooze, J. (1999). *Introduction to protein structure*. Garland.

[Brinkmann et al., 2005] Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., and Philippe, H. (2005). Indications from tree reconstruction artefacts in ancient phylogenies. *Syst. Biol.*, 16 :817–825.

[Broet et al., 2002] Broet, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *J. Comp. Biol.*, 9 :671–683.

[Bruno, 1996] Bruno, W. J. (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.*, 13(10) :1368–74.

[Bryant et al., 2005] Bryant, D., Galtier, N., and Poursat, M.-A. (2005). Likelihood calculations in molecular phylogenetics. In Gascuel, O., editor, *Mathematics of evolution and phylogeny*, pages 33–62. Oxford University Press.

[Bryngelson and Wolynes, 1987] Bryngelson, J. D. and Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, 84 :7524–7528.

[Buchete et al., 2004a] Buchete, N. V., Straub, J. E., and Thirumalai, D. (2004a). Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.*, 14 :225–232.

[Buchete et al., 2004b] Buchete, N. V., Straub, J. E., and Thirumalai, D. (2004b). Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci.*, 13 :862–874.

[Case et al., 2008] Case, D., A Darden, T. A., Cheatham III, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K. F., Paesani, F., Vanicek, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., and A, K. P. (2008). *AMBER 10*. University of California, San Francisco.

[Castresana, 2000] Castresana, J. (2000). Selection of conserved blocks from multiple alignment for their use in phylogenetic analysis. *Mol. Biol. Evol.*, 17 :540–552.

[Chen and Shakhnovich, 2005] Chen, W. W. and Shakhnovich (2005). Lessons from the design of a novel atomic potential for protein folding. *Proteins science*, 14 :1741–1752.

[Chhajer and Crippen, 2002] Chhajer, M. and Crippen, G. M. (2002). A protein folding potential that place the native states of a large number of proteins near a local optimum. *BMC Struct. Biol.*, 2 :4.

[Chib, 1995] Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Am. Stat. Assoc.*, 90 :1313–1321.

[Chib and Jeliazkov, 2001] Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Am. Stat. Assoc.*, 96 :270–281.

[Chiu and Goldstein, 1998a] Chiu, T. L. and Goldstein, R. A. (1998a). Optimized potentials for the inverse protein folding problem. *Protein Eng.*, 11 :749–752.

[Chiu and Goldstein, 1998b] Chiu, T. L. and Goldstein, R. A. (1998b). Optimizing energy potentials for success in protein tertiary structure prediction. *Fold. Des.*, 3 :223–228.

[Chiu and Goldstein, 2000] Chiu, T. L. and Goldstein, R. A. (2000). How to generate improved potentials for protein tertiary structure prediction : A lattice model study. *Proteins*, 41 :157–163.

[Choi et al., 2007] Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H., and Thorne, J. L. (2007). Quantifying the impact of protein tertiary structure on molecular evolution. *Mol. Biol. Evol.*, 24(8) :1769–1782.

[Choi et al., 2008] Choi, S. C., Redelings, B. D., and Thorne, J. L. (2008). Basing population genetic inferences and models of molecular evolution upon desired stationary distribution of dna or protein sequences. *Phil. trans. R. Soc. B.*, 363 :3931–3939.

[Choi et al., 2009] Choi, S. C., Stone, E. A., Kishino, H., and Thorne, J. L. (2009). Estimates of natural selection due to protein tertiary structure inform the ancestry of biallelic loci. *Gene*, 441 :45–52.

[Chothia and Lesk, 1986] Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5(4) :823–826.

[Colovos and Yeates, 1993] Colovos, C. and Yeates, T. O. (1993). Verification of protein structures : Patterns of non-bonded atomic interactions. *Protein Sci.*, 2 :1511–1519.

[Cootes et al., 2000] Cootes, A. P., Curmi, P. M. G., and Torda, A. E. (2000). Biased Monte Carlo optimization of protein sequences. *J. Chem. Phys.*, 113 :2489–2496.

[Cox, 1961] Cox, D. R. (1961). Test of separate families of hypotheses. *Proceedings of the 4th Berkeley Symposium (University of California press)*, 1 :105–123.

[Creighton, 1993] Creighton, T. E. (1993). *Proteins, Structures and molecular properties.* Freeman.

[Dahiyat and Mayo, 1997] Dahiyat, B. I. and Mayo, S. L. (1997). De novo protein design : fully automated sequence selection. *Science*, 278 :82–86.

[Dahiyat et al., 1997] Dahiyat, B. I., Sarisky, C. A., and Mayo, S. L. (1997). De novo protein design : towards fully automated sequence selection. *J. Mol. Biol.*, 273 :789–796.

[Dantzig et al., 1955] Dantzig, G., Orden, A., and Wolfe, P. (1955). The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific J. Math*, 5 :183–195.

[Dayhoff et al., 1978] Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In Dayhoff, M., editor, *Atlas of Protein Sequence and Structure*, pages 345–352. National Biomedical Research Foundation, Washington, DC.

[Dayhoff et al., 1972] Dayhoff, M. O., Eck, R. V., and Park, C. M. (1972). A model of evolutionary change in proteins. In Dayhoff, M. O., editor, *Atlas of protein sequence and structure*, volume 5, pages 88–89. National Biomedical research fundation.

[Dehouck et al., 2004] Dehouck, Y., Gilis, D., and Rooman, M. (2004). Database-derived potentials dependent on protein size for in silico folding and design. *Biophys. J.*, 87 :171–81.

[Dehouck et al., 2006] Dehouck, Y., Gilis, D., and Rooman, M. (2006). A new generation of statistical potentials for proteins. *Biophys. J.*, 90 :4010–4017.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, 39 :1–38.

[Desjarlais and Clarke, 1998] Desjarlais, J. R. and Clarke, N. D. (1998). Computer search algorithm in protein modification and design. *Curr. Opin. Struct. Biol.*, 8 :471–475.

[Desmet et al., 1992] Desmet, J., Maeyer, M. D., Hazes, B., and Lasters, I. (1992). The dead end elimination theorem and its use in protein side-chain positioning. *Nature*, 356 :539–542.

[Deutsch and Kurowski, 1996] Deutsch, J. M. and Kurowski, T. (1996). New algorithm for protein design. *Phys. Rev. Lett.*, 76(2) :323–326.

[Deutsch and Kurowski, 1997] Deutsch, J. M. and Kurowski, T. (1997). Design of force fields from data at finite temperature. *Phys. Rev. Lett.*, 56(4) :4553–4556.

[Dimmic et al., 2000] Dimmic, M. W., Mindell, D. P., and Goldstein, R. A. (2000). Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac. Symp. Biocomput.*, 5 :18–29.

[Dong et al., 2006] Dong, Q., Wang, X., and Lin, L. (2006). Novel knowledge-based mean force potential at the profile level. *BMC Bioinformatics*, 7 :324.

[Drexler, 1981] Drexler, K. E. (1981). Molecular engineering : an approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A.*, 78 :5275–5278.

[Dunbrack and Karplus, 1993] Dunbrack, R. and Karplus, M. (1993). Backbone-dependent rotamer library for proteins : application to side-chain prediction. *J. Mol. Biol.*, 230 :543–574.

[Efron and Tibshirani, 1986] Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, 1 :54–77.

[Escobar and West, 1995] Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, 90 :577–588.

[Eskin et al., 2001] Eskin, E., Grundy, W. N., and Singer, Y. (2001). Using mixtures of common ancestors for estimating the probabilities of discrete events in biological sequences. *Bioinformatics*, 17 :S65–S73.

[Fast et al., 2001] Fast, N. M., Kissinger, J. C., Roos, D. S., and Keeling, P. J. (2001). Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol. Biol. Evol.*, 18 :418–426.

[Felsenstein, 1978] Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 27 :401–410.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from dna sequences : a maximum likelihood approach. *Mol. Biol. Evol.*, 17 :368–376.

[Feng et al., 2007] Feng, Y., kloczkowiski, A., and Jernigan, R. L. (2007). Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins : structure, function and genetics*, 68 :57–66.

[Ferguson, 1973] Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Ann. Stat.*, 1 :209–230.

[Finkelstein, 1997] Finkelstein, A. V. (1997). Protein structure : what is it possible to predict now ? *Curr. Opin. Struct. Biol.*, 7 :60–71.

[Finkelstein et al., 1995a] Finkelstein, A. V., Badretdinov, A. Y., and M, G. A. (1995a). Why do protein architecture have a boltzmann-like statistics ? *Proteins*, 23 :142–150.

[Finkelstein et al., 1995b] Finkelstein, A. V., Gutin, A. M., and Badretdinov, A. Y. (1995b). Boltzmann-like statistics of protein architectures. origins and consequences. *Sub-Cell. Biochem.*, 24 :1–26.

[Fischer et al., 1996] Fischer, D., Rice, D., Bowie, J. U., and Eisenberg, D. (1996). Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J.*, 10 :126–136.

[Foster, 2004] Foster, P. G. (2004). Modelling compositional heterogeneity. *Syst. Biol.*, 53(3) :485–495.

[Furuichi and Koehl, 1998] Furuichi, E. and Koehl, P. (1998). Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins*, 31 :139–149.

[Gaut and Lewis, 1995] Gaut, B. S. and Lewis, P. O. (1995). Success of the maximum likelihood phylogeny inference in the four taxon case. *Mol. Biol. Evol.*, 12 :152–162.

[Gelfand and Ghosh, 1998] Gelfand, A. E. and Ghosh, S. K. (1998). Model choice : A minimum posterior predictive loss approach. *Biometrika*, 85(1) :1–11.

[Gelman, 1998] Gelman, A. (1998). Simulating normalizing constants : from importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13 :163–185.

[Gelman et al., 2004] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Chapman and Hall/CRC.

[Gelman et al., 1996] Gelman, A., Meng, X. L., and Stern, H. (1996). Posterior predicive assessment of model fitness via realised discrepancies. *Statistica Sinica*, 6 :733–807.

[Geyer, 1992] Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Stat. Sci.*, 7 :473–483.

[Geyer, 1994] Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. *Technical report, school of statistics, University of Minnesota*, 568.

[Gibbons et al., 2004] Gibbons, D. L., Vaney, M.-C., Roussel, A., Vigouroux, A., Reilly, B., Lepault, J., Kielian, M., and A, R. F. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Nature*, 427 :320–325.

[Glaser et al., 2005] Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T., and Ben-Tal, N. (2005). The consurf-hssp database : The mapping of evolutionary conservation among homologs onto pdb structures. *Proteins : Struct., Funct., and Bioinformat.*, 58 :610–617.

[Godzik et al., 1995] Godzik, A., Kolinski, A., and Skolnick, J. (1995). Are proteins ideal mixtures of amino acids ? analysis of energy parameter sets. *Protein Sci.*, 4 :2107–2117.

[Goldman, 1993] Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.*, 36(2) :182–198.

[Goldman et al., 1998] Goldman, N., Thorne, J., and Jones, D. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149 :445–458.

[Goldman et al., 1996] Goldman, N., Thorne, J. L., and T, J. D. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Mol. Biol. Evol.*, 263 :196–208.

[Goldman and Whelan, 2002] Goldman, N. and Whelan, S. (2002). A novel use of equilibrium frequencies in models of sequence evolution. *Mol. Biol. Evol.*, 19(11) :1821–1831.

[Goldman and Yang, 1994] Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol. Biol. Evol.*, 11 :725–736.

[Goldstein et al., 1992] Goldstein, R. A., Luthey-Schulten, Z. A., and Wolynes, P. G. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. USA*, 89 :4918–4922.

[Goldstein, 1994] Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, 66 :1335–1340.

[Gordon et al., 2003] Gordon, D. B., Hom, G. K., Mayo, S. L., and Pierce, N. A. (2003). Exact rotamer optimization for protein design. *J. Comput. Chem.*, 24 :232–243.

[Gordon and Mayo, 1998] Gordon, D. B. and Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead end elimination theorem. *J. Comput. Chem.*, 19 :1505–1514.

[Grantham, 1974] Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)*, 185(4154) :862–864.

[Green, 1995] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4) :711–732.

[Green and Richardson, 1998] Green, P. J. and Richardson, S. (1998). *Modelling heterogeneity with and without the Dirichlet process.* Technical report, University of Bristol.

[Gromiha and Selvaraj, 2004] Gromiha, M. M. and Selvaraj, S. (2004). Inter-residue interactions in protein folding and stability. *Progress in Biophys. and Mol. Biol.*, 86 :235–277.

[Guindon, 2003] Guindon, S. (2003). *Méthodes et algorithmes pour l'approche statistique en phylogénie.* PhD thesis, Université de Montpellier 2.

[Guindon and Gascuel, 2003] Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52(5) :696–704.

[Guindon et al., 2004] Guindon, S., Rodrigo, A., Dyer, K. A., and Huelsenbeck, J. P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl. Acad. Sci. USA*, 101(35) :12957–12962.

[Günter et al., 1997] Günter, P., Mumenthaler, C., and Wüthrich (1997). Torsion angle dynamics for nmr structure calculation with the new program dyana. *J. Mol. Biol.*, 273 :283–298.

[Halpern and Bruno, 1998] Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences : Modeling site-specific residue frequencies. *Mol. Biol. Evol.*, 15(7) :910–917.

[Han and Carlin, 2000] Han, C. and Carlin, B. P. (2000). MCMC methods for computing Bayes factors : a comparative review. *Biometrika*, 82(4) :711–732.

[Hastings, 1970] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57 :97–109.

[Hellinga, 1998] Hellinga, H. W. (1998). Computational protein engineering. *Nat. Struct. Biol.*, 5 :525–527.

[Hellinga and Richards, 1994] Hellinga, H. W. and Richards, F. M. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 91 :5803–5807.

[Hendlich et al., 1990] Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. the calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.*, 216 :167–180.

[Holder and Lewis, 2003] Holder, M. and Lewis, P. O. (2003). Phylogenetic estimation : traditional and Bayesian approaches. *Nat. Rev. Genet.*, 4 :275–284.

[Hubbard and Thornton, 1993] Hubbard, S. J. and Thornton, J. M. (1993). Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London.*

[Huelsenbeck and Imennov, 2002] Huelsenbeck, J. P. and Imennov, N. S. (2002). Geographic origin of human mitochondrial DNA : accomodating phylogenetic uncertainty and model comparison. *Syst. Biol.*, 51(1) :155–165.

[Huelsenbeck et al., 2004] Huelsenbeck, J. P., Larget, B., and Alfaro, M. E. (2004). Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.*, 21(6) :1123–1133.

[Huelsenbeck et al., 2002] Huelsenbeck, J. P., Larget, B., Miller, R. E., and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.*, 51(5) :673–688.

[Huelsenbeck and Nielsen, 1999] Huelsenbeck, J. P. and Nielsen, R. (1999). Variation in the pattern of nucleotide substitution across sites. *J. Mol. Evol.*, 48(1) :86–93.

[Huelsenbeck and Rannala, 1997] Huelsenbeck, J. P. and Rannala, B. (1997). Phylogenetic methods come of age : testing hypotheses in an evolutionary context. *Science*, 276(5310) :227–232.

[Huelsenbeck et al., 2000] Huelsenbeck, J. P., Rannala, B., and Masly, J. P. (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, 288(5475) :2349–2350.

[Huelsenbeck and Ronquist, 2001] Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES : Bayesian inference of phylogenetic trees. *Bioinformatics*, 17 :754–755.

[Huelsenbeck et al., 2001] Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550) :2310–2314.

[Irbäck et al., 1998] Irbäck, A., Peterson, C., Potthast, F., and Sandelin, E. (1998). Monte Carlo procedure for protein design. *Phys. Rev. E*, 58 :R5249–R5252.

[Irestedt et al., 2004] Irestedt, M., Fjeldsa, J., Nylander, J. A., and Ericson, P. G. (2004). Phylogenetic relationships of typical antbirds (Thamnophilidae) and test of incongruence based on Bayes factors. *BMC Evol. Biol.*, 4 :23.

[Jaramillo et al., 2002] Jaramillo, A., Wernisch, L., Héry, S., and Wodak, S. J. (2002). Folding free energy function selects native-like protein sequences in the core but not on the surface. *Proc. Natl. Acad. Sci. U.S.A.*, 99 :13554–13559.

[Jaynes, 2003] Jaynes, E. (2003). *Probability Theory. The logic of science*. Cambridge University Press.

[Jeffreys, 1935] Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc.*, 31 :203–222.

[Jeffreys, 1961] Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

[Jernigan and Bahar, 1996] Jernigan, R. L. and Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.*, 6 :195–209.

[Jiang et al., 2000] Jiang, X., Farid, H., Pistor, E., and Farid, R. S. (2000). A new approach to the design of uniquely folded thermally stable proteins. *Protein Sci.*, 9 :403–416.

[Jones et al., 1992a] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992a). A new approach to protein fold recognition. *Nature*, 358 :86–89.

[Jones et al., 1992b] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992b). The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8 :275–282.

[Jones et al., 1992c] Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992c). The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8 :275–282.

[Jones and Thornton, 1996] Jones, D. T. and Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.*, 6 :210–216.

[Jukes and Cantor, 1969] Jukes, T. H. and Cantor, C. R. (1969). *Evolution of Protein Molecules*. Academy Press.

[Kass and Raftery, 1995] Kass, R. and Raftery, A. (1995). Bayes factors and model uncertainty. *J. Am. Stat. Assoc.*, 90 :773–795.

[Kimura, 1968] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217 :624–626.

[Kimura, 1983] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

[Kimura and Ohta, 1974] Kimura, M. and Ohta, T. (1974). On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA*, 71(7) :2848–2852.

[Kleinman et al., 2006] Kleinman, C. L., Rodrigue, N., Bonnard, C., Philippe, H., and Lartillot, N. (2006). A maximum likelihood framework for protein design. *BMC Bioinformatics*, 7 :326–343.

[Kleinman et al., Submitted] Kleinman, C. L., Rodrigue, N., Philippe, H., and Lartillot, N. (2009). Protein structure representations for evolutionary analysis.

[Kocher et al., 1994] Kocher, J., Rooman, M., and Wodak, S. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.*, 235 :1598–1613.

[Koehl and Delarue, 1994] Koehl, P. and Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.*, 239 :249–275.

[Koehl and Delarue, 1996] Koehl, P. and Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.*, 6 :222–226.

[Koehl and Levitt, 1999a] Koehl, P. and Levitt, M. (1999a). De novo protein design. I. in search of stability and specificity. *J. Mol. Biol.*, 293 :1161–1181.

[Koehl and Levitt, 1999b] Koehl, P. and Levitt, M. (1999b). De novo protein design. II. plasticity in sequence space. *J. Mol. Biol.*, 293 :1183–1193.

[Koehl and Levitt, 2002] Koehl, P. and Levitt, M. (2002). Protein topology and stability define the space of allowed sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 99 :1280–1285.

[Kono and Saven, 2001] Kono, H. and Saven, J. G. (2001). 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.

[Koshi and Goldstein, 1997] Koshi, J. M. and Goldstein, R. A. (1997). Mutation matrices and physical-chemical properties : Correlations and implications. *Proteins*, 27 :336–344.

[Koshi and Goldstein, 1998] Koshi, J. M. and Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins*, 32 :289–295.

[Koshi and Goldstein, 2001] Koshi, J. M. and Goldstein, R. A. (2001). Analyzing site heterogeneity during protein evolution. *Pac. Symp. Biocomput.*, pages 191–202.

[Koshi et al., 1999] Koshi, J. M., Mindell, D. P., and Goldstein, R. A. (1999). Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. *Mol. Biol. Evol.*, 16 :173–179.

[Kosiol and Goldman, 2005] Kosiol, C. and Goldman, N. (2005). Different versions of the dayhoff rate matrix. *Mol. Biol. Evol.*, 22(2) :193–199.

[Kosiol et al., 2007] Kosiol, C., Holmes, I., and Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.*, 24(7) :1464–1479.

[Kuhlman and Baker, 2000] Kuhlman, B. and Baker, D. (2000). Native protein sequences are close to optimal for their structure. *Proc. Natl. Acad. Sci. USA*, 97(19) :10383–10388.

[Kuhner et al., 1995] Kuhner, M. K., Yamato, J., and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 149 :1421–1430.

[Kurochkina and Lee, 1995] Kurochkina, N. and Lee, B. (1995). Hydrophobic potential by pairwise surface area sum. *Protein Eng.*, 8 :437–442.

[Kurosky and Deutsch, 1995] Kurosky, T. and Deutsch, J. M. (1995). Design of copolymeric material. *J. Phys. A : Math. Gen.*, 27 :L387–L393.

[Kussel et al., 2002] Kussel, E., Shimada, J., and Shakhnovich, E. I. (2002). A structure-based method for derivation of all-atom potentials for protein folding. *Proc. Natl. Acad. Sci. USA*, 99(8) :5343–5348.

[Lanave et al., 1984] Lanave, C., Preparato, G., Sacone, C., and Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, 20(1) :86–93.

[Lang et al., 2002] Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W., and Burger, G. (2002). The closest unicellular relatives of animals. *Curr. Biol.*, 12 :1773–1778.

[Larget, 2005] Larget, B. (2005). Introduction to markov chain monte carlo methods in molecular evolution. In York, S. N., editor, *Statistics for Biology and Health*, pages 45–62. Springer.

[Larget and Simon, 1999] Larget, B. and Simon, D. L. (1999). Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.*, 116(6) :750–759.

[Larson et al., 2002] Larson, S. M., England, J. L., Desjarlais, J. R., and Pande, V. S. (2002). Thoroughly sampling sequence space : large-scale protein design of structural ensembles. *Protein Sci.*, 11 :2084–2813.

[Lartillot et al., 2009] Lartillot, N., Lepage, T., and Blanquart, S. (2009). Phylobayes 3. a bayesian software for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17) :2286–2288.

[Lartillot and Philippe, 2004] Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid remplacement process. *Mol. Biol. Evol.*, 21(6) :1095–1109.

[Lartillot and Philippe, 2006] Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55 :195–207.

[Laskowski et al., 2005] Laskowski, R. A., Chistyakov, V. V., and M., T. J. (2005). Pdbsum more : new summaries and analyses of the known 3d structures of proteins and nucleic acids. *Nucleic Acids Res.*, 33 :D266–D268.

[Launay et al., 2007] Launay, G., Mendez, R., Wodak, S., and Simonson, T. (2007). Recognizing protein-protein interfaces with empirical potentials and reduced amino acid alphabets. *BMC Bioinformatics*, 8 :270.

[Lazaridis and Karplus, 2000] Lazaridis, T. and Karplus, M. (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.*, 10 :139–145.

[Le and Gascuel, 2008] Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, 25(7) :1307–1320.

[Le et al., 2008a] Le, S. Q., Gascuel, O., and Lartillot, N. (2008a). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 29 :2317–2323.

[Le et al., 2008b] Le, S. Q., Lartillot, N., and Gascuel, O. (2008b). Phylogenetic mixture models for proteins. *Phil. trans. of the Royal society B.*, 363 :3965–3976.

[Lee and Richards, 1971] Lee, B. and Richards, F. M. (1971). The interpretation of protein structures : estimation of static accessibility. *J. Mol. Biol.*, 55 :379–400.

[Li, 1996] Li, S. (1996). *Phylogenetic tree construction using Markov chain Monte Carlo.* Phd dissertation, Ohio State University, Columbus.

[Lìo and Goldman, 1999] Lìo, P. and Goldman, N. (1999). Using protein structural information in evolutionary inference : transmembrane proteins. *Mol. Biol. Evol.*, 16 :1696–1710.

[Livingstone and Barton, 1993] Livingstone, C. D. and Barton, G. J. (1993). Protein sequence alignments : A strategy for the hierarchical analysis of residue conservation. *Comp. Appl. Bio. Sci*, 9 :745–756.

[Liwo et al., 2004] Liwo, A., Oldziej, S., Czaplewski, C., Urszula, K., and Scheraga, H. A. (2004). Parametrization of backbon-electrostatic and multibody contributions to the unres force field for protein-structure prediction from ab-initio energy surfaces of model systems. *J. Phys. Chem.*, 108 :9421–9438.

[Looger and Hellinga, 2001] Looger, L. and Hellinga, H. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable : implications for protein design and structural genomics. *J. Mol. Biol.*, 307 :429–445.

[Loose et al., 2004] Loose, C., Klepeis, J. L., and Floudas, C. A. (2004). A new pairwise folding potential based on improved decoy generation and side-chain packing. *Proteins*, 54 :303–314.

[Lopez, 1997] Lopez, P. (1997). *Analyse phylogenetique de grands alignements de proteines : vers une classification des sites ?* Master degree dissertation, Universite Paris XI.

[MacKerel Jr et al., 1998] MacKerel Jr, A. D., Brooks III, C. L., Nilsson, L., Roux, B., Won, Y., and Karplus, M. (1998). Charmm : The energy function and its parameterization with an overview of the program. In v. R. Schleyer et al., P., editor, *The Encyclopedia of Computational Chemistry*, volume 1, pages 271–277. John Wiley & Sons : Chichester.

[Maiorov and Crippen, 1992] Maiorov, V. N. and Crippen, G. M. (1992). Contact potential that recognozes the correct folding of globular proteins. *J. Mol. Biol.*, 227 :876–888.

[Maupetit et al., 2007] Maupetit, J., Tuffery, P., and P, D. (2007). A coarse-grained protein force field for folding and structure prediction. *Proteins*, 69 :394–408.

[Meller and Elber, 2001] Meller, J. and Elber, R. (2001). Linear optimization and a double statistical filter for protein threading protocols. *Proteins*, 45 :241–261.

[Mendes et al., 2002] Mendes, J., Guerois, R., and Serrano, L. (2002). Energy estimation in protein design. *Curr. Opin. Struct. Biol.*, 12(4) :441–446.

[Meng, 1994] Meng, X. L. (1994). Posterior predictive p-values. *Ann. Stat.*, 22 :1142–1160.

[Meng and Wong, 1996] Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalising constants via a simple identity : a theoretical exploration. *Statistica Sinica*, 6 :831–860.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21 :1087–1092.

[Meyerguz et al., 2004] Meyerguz, L., Grasso, C., Kleinberg, J., and Elber, R. (2004). Computational analysis of sequence selection mechanisms. *Structure*, 12 :547–557.

[Micheletti et al., 1999] Micheletti, C., Maritan, A., and Banavar, J. R. (1999). A comparative study of existing and new design techniques for protein models. *J. Chem. Phys.*, 110 :9730–9738.

[Micheletti et al., 2001] Micheletti, C., Seno, F., Banavar, J. R., and Maritan, A. (2001). Learning effective amino acid interactions through iterative stochastic techniques. *Proteins*, 42 :422–431.

[Micheletti et al., 1998] Micheletti, C., Seno, F., Maritan, A., and Banavar, J. (1998). Design of proteins with hydrophobic and polar amino acids. *Proteins*, 32 :80–87.

[Minin et al., 2003] Minin, V., Abdo, Z., Joyce, P., and Sullivan, J. (2003). Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.*, 52 :674–683.

[Mirny and Shakhnovich, 1996] Mirny, L. A. and Shakhnovich, E. I. (1996). How to derive a protein folding potential ? a new approach to an old problem. *J. Mol. Evol.*, 264 :1164–1179.

[Miyamoto and Fitch, 1996] Miyamoto, M. M. and Fitch, W. M. (1996). Constraints on protein evolution and the age of eubacteria/eukaryote split. *Syst. Biol.*, 45 :568–575.

[Miyazawa and Jernigan, 1985] Miyazawa, S. and Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures : quasi-chemical approximation. *Macromolecules*, 18 :534–552.

[Miyazawa and Jernigan, 1996] Miyazawa, S. and Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable hight packing density term, for simulation and threading. *J. Mol. Biol.*, 256(3) :623–644.

[Momany et al., 1975] Momany, F. A., McGuire, R. F., Burgess, A. W., and Scheraga, H. A. (1975). Energy parameters in polypeptides. vii. geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions and intriseict tortional potentials for the naturally occuring amio acids. *J. Phys. Chem.*, 79 :2361–2381.

[Moore and Maranas, 2003] Moore, G. L. and Maranas, C. D. (2003). Identifying residue-residue clashes in protein hybrids by using a second-order mean-field approach. *Proc. Natl. Acad. Sci. U.S.A.*, 100 :5091–5096.

[Moretti et al., 2006] Moretti, S., Reinier, F., Poirot, O., Armougom, F., Audic, S., Keduas, V., and Notredame, C. (2006). Protogene : turning amino acid alignments into bona fide cds nucleotide alignments. *Nucleic Acid Res.*, 34 :W600–W603.

[Morrissey and Shakhnovich, 1996] Morrissey, M. P. and Shakhnovich, E. I. (1996). Design of proteins with selected thermal properties. *Fold. Des.*, 1 :391–405.

[Moult, 1997] Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.*, 7(2) :1994–1999.

[Muller et al., 2002] Muller, T., Spang, R., and Vingron, M. (2002). Estimating amino acid substitution models : a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol. Biol. Evol.*, 19 :8–13.

[Murphy et al., 2001] Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W., and Springer, M. S. (2001). Resolution of the early placental mammal radiation using bayesian phylogenics. *Science*, 294(5550) :2348–2351.

[Muse and Gaut, 1994] Muse, S. V. and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions, with applications to cholorplast genome. *Mol. Biol. Evol.*, 11 :715–724.

[Neal, 1993] Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods. Technical report, University of Toronto.

[Neal, 2000] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.*, 9 :249–265.

[Newton and Raftery, 1994] Newton, M. A. and Raftery, A. E. (1994). Approximating Bayesian inference with the weigthed likelihood bootstrap. *J. R. Stat. Soc. B*, 56 :3–48.

[Nielsen, 2001] Nielsen, R. (2001). Mapping mutations on phylogenies. *Syst. Biol.*, 51 :729–739.

[Nielsen and Huelsenbeck, 2002] Nielsen, R. and Huelsenbeck, J. P. (2002). Detecting positively selected amino acid sites using posterior predictive p-values. *Pac. Symp. Biocomput.*, pages 576–588.

[Nielsen and Yang, 1998] Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics*, 148 :929–936.

[Noirel, 2006] Noirel, J. (2006). *Évolution in silico des protéines monomériques et dimériques.* PhD thesis, École doctorale de l'École Polytechnique.

[Nylander et al., 2004] Nylander, J. A. A., Ronquist, F., Huelsenbeck, J. P., and Nieves-Aldrey, J. L. (2004). Bayesian phylogenetic analysis of combined data. *Syst. Biol.*, 53(1) :47–67.

[Ogata, 1989] Ogata, Y. (1989). A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, 55 :137–157.

[O'Hagan, 1995] O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *J. R. Stat. Soc. B*, 57 :99–138.

[Ohta, 1973] Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246 :96–98.

[Pabo, 1983] Pabo, C. (1983). Molecular technology : designing proteins and peptides. *Nature*, 301 :200.

[Pagel and Meade, 2004] Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, 53 :561–581.

[Pande et al., 1997] Pande, V. S., Grosberg, A. Y., and Tanaka, T. (1997). Statistical mechanics of simple model of protein folding and design. *Biophys. J.*, 73(6) :3192–3210.

[Panjkovich et al., 2008] Panjkovich, A., Melo, F., and Marti-Renom, M. (2008). Evolutionary potentials : structure specific knowledge-based potentials exploiting the evolutionary record of sequence homologs. *Genome Biol.*, 9.

[Parisi and Echave, 2001] Parisi, G. and Echave, J. (2001). Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.*, 18(5) :750–756.

[Park and Levitt, 1996] Park, B. and Levitt, M. (1996). Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.*, 258 :367–392.

[Park et al., 2000] Park, K., Vendruscolo, M., and Domany, E. (2000). Toward an energy function for the contact map representation of proteins. *Proteins*, 40 :237–248.

[Park et al., 2004] Park, S., Yang, X., and Saven, J. G. (2004). Advances in computational protein design. *Curr. Opin. Struct. Biol.*, 14 :487–494.

[Philippe, 1993] Philippe, H. (1993). MUST, a computer package of management utilities for sequences and trees. *Nucleic Acid Res.*, 21 :5264–5272.

[Philippe et al., 2005] Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecysozoa, Lophotrochozoa and Protostomia. *Mol. Biol. Evol.*, 22 :1246–1253.

[Ponders and Richards, 1987] Ponders, J. W. and Richards, F. M. (1987). Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, 193 :775–791.

[Posada and Crandall, 2001] Posada, D. and Crandall, K. (2001). Selecting the best-fit model of nucleotide substitution. *Syst. Biol.*, 50 :580–601.

[Pushkarev et al., 2009] Pushkarev, D., Neff, N. F., and Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nature Biotech.*

[Qiu and Elber, 2005] Qiu, J. and Elber, R. (2005). Atomically detailed potentials to recognize native and approximate protein structures. *Proteins*, 61 :44–55.

[Raftery and Lewis, 1992] Raftery, A. E. and Lewis, S. M. (1992). [practical Markov chain Monte Carlo] : Comment : one long run with diagnostics : implementation strategies for Markov chain Monte Carlo. *Stat. Sci.*, 7 :493–497.

[Rajgaria et al., 2008] Rajgaria, R., McAllister, S. R., and Floudas, C. A. (2008). Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins*, 70(3) :950–970.

[Rannala, 2002] Rannala, B. (2002). Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.*, 51 :754–760.

[Robinson et al., 2003] Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.*, 20(10) :1692–1704.

[Rodrigue, 2007] Rodrigue, N. (2007). *Phylogenetic structural modeling of molecular evolution.* PhD thesis, Université de Montréal.

[Rodrigue et al., 2009] Rodrigue, N., Kleinman, C. L., Philippe, H., and Lartillot, N. (2009). Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol. Biol. Evol.*, 26(7) :1663–1676.

[Rodrigue et al., 2005] Rodrigue, N., Lartillot, N., Bryant, D., and Philippe, H. (2005). Site interdependence attributed to tertiary structure in protein evolution. *Gene*, 347(2) :207–217.

[Rodrigue et al., 2008a] Rodrigue, N., Lartillot, N., and Philippe, H. (2008a). Bayesian comparisons of codon substitution models. *Genetics*, 180 :1579–1591.

[Rodrigue et al., 2006] Rodrigue, N., Philippe, H., and Lartillot, N. (2006). Assessing site-interdependent phylogenetic models of sequence evolution. *Mol. Biol. Evol.*, 23(9) :1762–1775.

[Rodrigue et al., 2007] Rodrigue, N., Philippe, H., and Lartillot, N. (2007). Exploring fast computational strategies for probabilistic phylogenetic analysis. *Syst. Biol.*, 56(5) :711–726.

[Rodrigue et al., 2008b] Rodrigue, N., Philippe, H., and Lartillot, N. (2008b). Uniformization for sampling realisations of markov processes : applications to bayesian implementations of codon substitution models. *Bioinformatics*, 24(1) :56–62.

[Rojnuckarin and Subramaniam, 1999] Rojnuckarin, A. and Subramaniam, S. (1999). Knowledge-based interaction potentials for proteins. *Proteins*, 36 :54–67.

[Rooman and Gilis, 1998] Rooman, M. and Gilis, D. (1998). Different derivations of knowledge-based potentials and analysis of their robustness and context-dependent predictive power. *Eur. J. Biochem.*, 254 :135–143.

[Ross and Rodrigo, 2002] Ross, H. A. and Rodrigo, A. (2002). Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J. virol.*, 76(22) :11715–11720.

[Rossi et al., 2000] Rossi, A., Maritan, A., and Micheletti, C. (2000). A novel iterative strategy for protein design. *J. Chem. Phys.*, 112(4) :2050–2055.

[Rossi et al., 2001] Rossi, A., Micheletti, C., Seno, F., and Maritan, A. (2001). A self-consistent knowledge-based approach to protein design. *Biophys. J.*, 80 :480–490.

[Rubin, 1984] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.*, 4 :1151–1172.

[Saitou and M, 1987] Saitou, N. and M, N. (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4 :406–425.

[Samudrala and Levitt, 2000] Samudrala, R. and Levitt, M. (2000). Decoys 'r' us : A database of incorrect protein conformations to improve protein structure prediction. *Protein Sci.*, 9 :1399–1401.

[Samudrala et al., 1999] Samudrala, R., Xia, Y., Levitt, M., and Huang, E. S. (1999). A combined approach for ab initio construction of low resolution protein teritary structures from sequences. *Pac. Symp. Biocomput.*, 4 :505–516.

[Saven and Wolynes, 1997] Saven, J. G. and Wolynes, P. G. (1997). Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J. Phys. Chem.*, 101 :8375–8389.

[Schneider and Stephens, 1990] Schneider, T. D. and Stephens, R. M. (1990). Sequence logos : a new way to display consensus sequences. *Nucleic Acid Res.*, 18 :6097–6100.

[Schwartz, 1978] Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6(2) :461–464.

[Scott et al., 1997] Scott, W. R. P., Hunenberger, P. H., Trioni, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Kruger, P., and VanGunsteren, W. F. (1997). The gromos biomolecular simulation program package. *J. Phys. Chem.*, 103 :3596–3607.

[Seno et al., 1998] Seno, F., Micheletti, M., Maritan, A., and Banavar, J. R. (1998). Variational approach to protein design and extraction of interactional potentials. *Phys. Rev. Lett.*, 81 :2172–2175.

[Seno et al., 1996] Seno, F., Vendruscolo, M., Maritan, A., and Banavar, J. R. (1996). Optimal protein design procedures. *Phys. Rev. Lett.*, 77(9) :1901–1904.

[Shakhnovich and Gutin, 1993] Shakhnovich, E. and Gutin, A. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA*, 90(15) :7195–7199.

[Shirota et al., 2008] Shirota, M., Ishida, T., and Kinoshita, K. (2008). Effects of surface-to-volume ratio of proteins on hydrophilic residues : Decrease in occurrence and increase in buried fraction. *Protein Sci.*, 17(9) :1596–1602.

[Simon et al., 1997] Simon, K. T., Kooperberg, C., Huang, C., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring function. *J. Mol. Biol.*, 268 :209–225.

[Sippl, 1990] Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213 :859–883.

[Sippl, 1993a] Sippl, M. J. (1993a). Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computationnal determination of protein structures. *J. Comput. Aided Mol. Des.*, 7 :473–501.

[Sippl, 1993b] Sippl, M. J. (1993b). Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17 :355–362.

[Sippl, 1995] Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Biol.*, 5 :229–235.

[Skolnick et al., 1997] Skolnick, J., Jaroszewski, L., Kolinski, A., and Godzik, A. (1997). Derivation and testing of pair potentials for protein folding. when is the quasichemical approximation correct ? *Protein Sci.*, 6(3) :676–688.

[Sohl et al., 1998] Sohl, J. L., Jaswal, S. S., and Agard, D. A. (1998). Unfolded conformations of α-lytic protease are more stable than its native state. *Nature*, 395 :817–819.

[Solis and Rackovsky, 2006] Solis, A. D. and Rackovsky, S. (2006). Improvement of statistical potentials and threading score functions using information maximixation. *Proteins*, 62(4) :892–908.

[Soyer et al., 2002] Soyer, O., Dimmic, M., Neubig, R., and Goldstein, R. (2002). Using evolutionary methods to study G-protein coupled receptors. *Pac. Symp. Biocomput.*, pages 625–636.

[Stefanovic et al., 2004] Stefanovic, S., Rice, D., and Palmer, J. (2004). Long branch attraction, taxon sampling, and the earliest angiosperms : Amborella or monocots ? *BMC Evol. Biol.*, 4 :35.

[Stone and Sidow, 2005] Stone, E. A. and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.*, 15 :978–986.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B*, 36 :111–147.

[Suchard et al., 2003] Suchard, M., Kitchen, C. M. R., Sinsheimer, J., and Weiss, R. E. (2003). Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst. Biol.*, 52 :649–664.

[Suchard et al., 2001] Suchard, M., Weiss, R., and Sinsheimer, J. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.*, 18 :1001–1013.

[Sullivan and Swofford, 1997] Sullivan, J. and Swofford, D. L. (1997). Are guinea pigs rodents ? the importance of adequate models in molecular phylogenetics. *J. Mammal. Evol.*, 4 :77–86.

[Sullivan and Swofford, 2001] Sullivan, J. and Swofford, D. L. (2001). Should we used model-based methods for phylogenetic inference when we know that assumptions about among-site variation and nucleotide substitution pattern are violated ? *Syst. Biol.*, 50 :723–729.

[Sun et al., 1995] Sun, S., Brem, R., Chan, R., and Dill, K. (1995). Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.*, 8(12) :1205–1213.

[Swofford, 1993] Swofford, D. (1993). PAUP : phylogenetic analysis using parsimony, version 3.1.1.

[Swofford et al., 1996] Swofford, D., Olsen, G., Waddell, P., and Hillis, D. (1996). *Phylogenetic inference, in Molecular Systematics*. Sinauer Associates.

[Tanaka and Scheraga, 1976] Tanaka, S. and Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structure of proteins. *Macromolecules*, 9(6) :945–950.

[Taverna and Goldstein, 2002] Taverna, D. M. and Goldstein, R. A. (2002). Why are proteins so robust to site mutations ? *J. Mol. Biol.*, 315 :479–484.

[Thomas and Dill, 1996a] Thomas, P. D. and Dill, K. A. (1996a). An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. USA*, 93 :11628–11633.

[Thomas and Dill, 1996b] Thomas, P. D. and Dill, K. A. (1996b). Statistical potentials extracted from protein structures : how accurate are they ? *J. Mol. Evol.*, 257 :457–469.

[Thompson et al., 1994] Thompson, J., Higgins, D., and Gibson, T. (1994). CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22 :4673–4680.

[Thorne et al., 2007] Thorne, J. L., Choi, S. C., Yu, J., Higgs, P. G., and Kishino, H. (2007). Population genetics without intraspecific data. *Mol. Biol. Evol.*, 24(8) :1667–1677.

[Thorne et al., 1996] Thorne, J. L., Goldman, N., and Jones, D. T. (1996). Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, 13(5) :666–673.

[Tiana et al., 2004] Tiana, G., Colombo, M., Provasi, D., and Broglia, R. A. (2004). Deriving amino acid contact potentials from their frequencies of occurrence in proteins : a lattice model study. *J. Phys. : Condens. Matter*, 16 :2551–2564.

[Tobi and Elber, 2000] Tobi, D. and Elber, R. (2000). Distance-dependent, pair potential for protein folding : results from linear optimization. *Proteins*, 41 :40–46.

[Tobi et al., 2000] Tobi, D., Shafran, G., Linial, N., and Elber, R. (2000). On the design and analysis of protein folding potentials. *Proteins*, 40 :71–85.

[Tozzini, 2005] Tozzini, V. (2005). Coarse-grained model for proteins. *Curr. Opin. Struct. Biol.*, 15 :144–150.

[Vendruscolo and Domany, 1998a] Vendruscolo, M. and Domany, E. (1998a). Efficient dynamics in the space of contact maps. *Fold. Des.*, 3 :329–338.

[Vendruscolo and Domany, 1998b] Vendruscolo, M. and Domany, E. (1998b). Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.*, 109(24) :11101–11108.

[Vendruscolo et al., 1997] Vendruscolo, M., Kussel, E., and Domany, E. (1997). Recovery of protein structure from contact maps. *Fold. Des.*, 2 :295–306.

[Vendruscolo et al., 2000] Vendruscolo, M., Najmanovich, R., and Domany, E. (2000). Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading. *Proteins*, 38 :134–148.

[Verdinelli and Wasserman, 1995] Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Am. Stat. Assoc.*, 90 :614–618.

[Waddell et al., 2002] Waddell, P. J., Kishino, H., and Ota, R. (2002). Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. *Genome Inform.*, 13 :82–92.

[Wald, 1949] Wald, A. (1949). Note on the consistency of maximumm likelihood. *Ann. Math. Stat.*, 20 :595–601.

[Wang and Dunbrack, 2003] Wang, G. and Dunbrack, R. L. J. (2003). Pisces : a protein sequence culling server. *Bioinformatics*, 19(12) :1589–1591.

[Wang et al., 2008] Wang, H. C., Li, K., Susko, E., and Roger, A. J. (2008). A class frequency mixture model that adjusts for the site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.*, 8 :331.

[Wernisch et al., 2000] Wernisch, L., Hery, S., and Wodak, S. J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.*, 301 :713–736.

[Whelan and Goldman, 2001] Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, 18 :691–699.

[Whelan and Goldman, 2004] Whelan, S. and Goldman, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, 167 :2027–2043.

[Whelan et al., 2001] Whelan, S., Lio, P., and Goldman, N. (2001). Molecular phylogenetics : state-of-the-art methods for looking into the past. *Trends. Genet.*, 17 :262–272.

[Williams et al., 2006] Williams, P. D., Pollock, D. D., Blackburne, B. P., and Goldstein, R. A. (2006). Assessing the accuracy of ancestral protein reconstruction methods. *PLOS computational biology*, 2(6) :598–605.

[Worth et al., 2009] Worth, C. L., Gong, S., and Blundell, T. L. (2009). Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell. Biol.*, 5(4) :823–826.

[Yang, 1993] Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10 :1396–1401.

[Yang, 1994] Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites : approximate methods. *J. Mol. Evol.*, 39 :306–14.

[Yang, 1995] Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics*, 139 :993–1005.

[Yang, 1996] Yang, Z. (1996). Among site variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.*, 11 :367–370.

[Yang and Bielawski, 2000] Yang, Z. and Bielawski, J. P. (2000). Statistical method for detecting molecular adaptation. *Trends Ecol. Evol.*, 15 :496–503.

[Yang and Nielsen, 2008] Yang, Z. and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.*, 25(3) :568–579.

[Yang et al., 2000a] Yang, Z., Nielsen, R., Goldman, N., and K, P. (2000a). Codon substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155 :431–449.

[Yang and Rannala, 1997] Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences : a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, 14 :717–724.

[Yang et al., 2000b] Yang, Z., Swanson, W. J., and Vacquier, V. D. (2000b). Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.*, 17 :1446–1455.

[Yu and Thorne, 2006] Yu, J. and Thorne, J. L. (2006). Dependance among sites in rna evolution. *Mol. Biol. Evol.*, 23(8) :1525–1537.

[Yue et al., 1995] Yue, K., Fiebig, K., Thomas, K. E., Chan, H. S., Shakhnovich, E. I., and Dill, K. A. (1995). A test of lattice protein folding algorithms. *Proc. Natl. Acad. Sci. USA*, 92 :325–329.

[Zhang and Eisenberg, 1994] Zhang, K. Y. J. and Eisenberg, D. (1994). The three-dimensional profile method using residue preference as a continuous function of residue environment. *Protein Sci.*, 3 :687–695.

[Zou and Saven, 2000] Zou, J. and Saven, J. G. (2000). Statistical theory for combinatorial libraries of folding proteins : energetic discrimination of a target structure. *J. Mol. Biol.*, 296 :281–294.