Prédiction de la structure des protéines.

M2 IGF

Julien Maupetit

Décembre 2007

Malgré des progrès importants, en 2007, la prédiction de la structure des protéines à partir de leur seule séquence en acides aminés reste encore un défi en bioinformatique structurale. A partir de la structure primaire d'une protéine, l'enjeu est de pouvoir déterminer, sans notion d'ordre, (i) les structures secondaires, *i.e.* les régions en hélices α ou en brins β, (ii) toutes les paires de brins β qui forment des feuillets (la topologie du feuillet β), (iii) les ponts disulfure si des cystéines sont présentes, (iv) les boucles qui connectent les structures secondaires, et enfin, (v) la structure tertiaire (repliement tridimensionnel) de la protéine.

Le problème de la prédiction de la structure des protéines a attiré un grand nombre de chercheurs venant de disciplines différentes, proposant des explications diverses à la question du repliement protéique. Selon l'hypothèse thermodynamique d'Anfinsen (1973), l'assemblage des protéines dans leur structure native n'est pas un processus biologique, mais purement physique, dépendant uniquement de la spécificité de sa séquence en acides aminés, et du solvant environnant. Cette hypothèse est aujourd'hui à nuancer, car nous savons d'une part qu'il existe un certain nombre de protéines chaperonnes aidant au repliement des protéines (Ellis, 2000, 2006) et, d'autre part, qu'il existe un certain nombre de protéines dites globalement désordonnées (Dyson and Wright, 2005), *i.e.* n'ayant pas de structures secondaires ni tertiaire bien déterminées, mais étant fonctionelles.

Levinthal (1968) s'est posé la question de savoir si ce phénomène spontanné était uniquement le fruit du hasard. Si l'on considère que chaque résidu peut adopter trois conformations distinctes (helice, brin ou coil), et que le temps de conversion d'une conformation à l'autre est de 10^{-13} secondes, alors, pour une protéine de 100 résidus, le temps moyen nécéssaire pour adopter une conformation serait de $3^{100} * 10^{-13} = 5.10^{-4}$ secondes, soit $1,6.10^{27}$ années ! Nous pouvons donc en conclure que le repliement des protéines ne se fait pas au hasard, et qu'il existe bien des forces capables de guider le repliement d'une protéine vers une conformation native dynamiquement la plus stable. Levinthal va même plus loin en émettant l'hypothèse qu'il doit exister des intermédiaires stables de repliement.

Cinq modèles principaux ont été proposés pour le repliement des protéines (Haspel et al., 2003) :

i. le modèle de nucléation-condensation, dans lequel une première étape de nucléation est suivie par une propagation rapide de la structure (Zimm and Bragg, 1959; Wetlaufer, 1973).

ii. le modèle de diffusion-collision, ici la nucléation intervient en différent points de la chaîne polypeptidique, ces noyaux diffusent puis s'unissent pour former des microstructures natives (Karplus and Weaver, 1994).

iii. le repliement séquentiel, pour lequel plusieurs segments de structure sont formés et assemblés à différents niveaux en suivant un chemin unique de repliement (Baldwin, 1975).

iv. le modèle d'effondrement hydrophobe, ce dernier implique que le repliement est directement guidé par les interactions hydrophobes formant le coeur protéique. Les structures secondaires sont formées dans un deuxième temps (Levitt and Warshel, 1975).

v. le modèle de repliement hiérarchique, pour lequel, le phénomène de nucléation est suivi par la formation des super-structures secondaires puis des domaines (Schulz, 1977; Lesk and Rose, 1981; Baldwin and Rose, 1999).

Chacun de ces modèles est corroboré par des analyses expérimentales (Tsai et al., 2002). Le repliement des protéines semble donc être un *continuum* entre ces différents modèles (Fersht and Daggett, 2002). Cependant, le modèle de repliement hiérarchique serait le plus communément admis (Haspel et al., 2003; Floudas, 2007). Il implique que des éléments des structures secondaires se forment rapidement, suivis par un réarrangement tridimensionnel plus lent pour former la structure tertiaire.

Dans la suite de cette introduction à la prédiction de la structure des protéines, nous allons commencer par décrire brièvement quelles méthodes ont été développées pour prédire leur structure locale, et ce, à différents niveaux d'organisation. Dans un deuxième temps, nous allons voir comment la prédiction de la structure tertiaire des protéines est envisagée, et quelles sont les méthodes les plus en vogues à l'heure actuelle. Enfin, nous présenterons les outils et serveurs disponibles pour réaliser une prédiction.

J'attire votre attention sur le fait que ce domaine est très en vogue et évolue très vite. De ce fait, la liste non exhaustive des méthodes présentées, valable à l'heure actuelle, est succeptible d'évoluer.

Première partie

Prédiction de la structure locale

1 Prédiction des structures secondaires

1.1 Avancées du domaine

La prédiction des structures secondaires des protéines à partir de leur seule séquence constitue un vaste domaine d'étude depuis les années 70.

Les premières méthodes publiées utilisent le fait que, dans les structures tridimensionnelles, certains acides aminés sont retrouvés préférentiellement dans certaines classes de structures secondaires (Chou and Fasman, 1974; Robson and Suzuki, 1976). Les préférences établies sont combinées par des règles de prédiction qui tiennent en compte les prédictions obtenues pour les sites voisins. Le taux de bonne prédiction de ces méthodes, noté Q_3 est de l'ordre de 55 à 58%. Le Q_3 correspond au pourcentage de bonne prédiction à trois états que sont les hélices α, feuillets β et coils.

Les méthodes développées dans les années 80, et jusqu'au début des années 90 tiennent comptent, dans le processus de prédiction, de l'environnement local des résidus dans la séquence. Ainsi, la prédiction de la conformation d'un résidu est réalisée à partir d'une fenêtre glissante dans la séquence. La taille de cette fenêtre est variable selon les méthodes et peut aller de 3 à 51 résidus (Rost, 2003).

La croissance des banques de données de séquences a permis de pouvoir utiliser l'information des séquences homologues. Les protéines homologues ayant des structures 3D similaires, elles ont de ce fait des structures secondaires similaires, et apportent donc une information utilisable par les méthodes de prédiction. Cette information est le plus souvent intégrée dans un profil, *i.e.* un tableau contenant les fréquences des acides aminés présents dans les colonnes d'un alignement multiple de séquences. L'intégration de ce type d'information dans les méthodes de prédiction permet d'atteindre des valeurs de Q_3 supérieures à 75%.

1.2 Les algorithmes

Une grande variété d'algorithmes a été appliquée au problème de la prédiction des structures secondaires.

La théorie de l'information est à la base des méthodes GOR (Garnier, Osguthorpe et Robson) (Garnier et al., 1978; Gibrat et al., 1987; Garnier et al., 1996) : elle permet de formuler l'influence de la séquence locale sur la conformation des résidus de manière rigoureuse.

Un grand nombre de méthodes utilisent les réseaux de neurones depuis les travaux précurseurs de Qian and Sejnowski (1988). La méthode PHD de Rost and Sander (1994) utilise une fenêtre de 13 résidus dans la séquence, et une prédiction en trois étapes : (i) un premier réseau de neurones (*sequence-to-structure*) prédit la structure secondaire du résidu central, (ii) un deuxième réseau (*structure-to-structure*) reçoit en entrée une fenêtre de 17 positions dans la prédiction fournie par le premier réseau et renvoie la structure secondaire du résidu central, enfin (iii) la moyenne est réalisée sur plusieurs réseaux de neurones (*jury decision*). L'information d'entrée est constituée par les profils issus des alignements multiples de la base HSSP (Sander and Schneider, 1991), ainsi que les poids relatifs à la conservation dans les colonnes de l'alignement.

La méthode PSIPRED de Jones (1999b) utilise une fenêtre de taille 15 et effectue également une correction des résultats du réseau par un deuxième réseau *structure-to-structure*.

L'information des séquences homologues est prise en compte sous la forme de profils générés par PSI-BLAST Altschul et al. (1997).

SSPRO développée par Baldi et al utilise des réseaux de neurones particuliers appelés réseaux bidirectionnels récurrents (Baldi et al., 1999; Pollastri et al., 2002). Riis and Krogh (1996) spécifient quant à eux des réseaux pour chaque classe structurale à prédire.

Les méthodes de type plus proches voisins utilisent les structures de protéines disponibles pour inférer la structure secondaire par comparaison de fragments (Levin et al., 1986; Salamov and Solovyev, 1997; Figureau et al., 2003; Jiang, 2003; Kim, 2004). La méthode SIMPA (Levin et al., 1986; Levin, 1997) utilise les scores d'alignements locaux pour effectuer la prédiction. Dans la méthode SOPMA (Geourjon and Deleage, 1995), les paramètres de prédiction sont optimisés pour la protéine à prédire, en utilisant des structures connues similaires. Ces deux méthodes appliquent un algorithme de régularisation des prédictions (Zimmermann, 1994).

Plus récemment, des méthodes basées sur les SVM (*Support Vector Machine*) ont été proposées (Hua and Sun, 2001; Kim and Park, 2003; Ward et al., 2003; Hu et al., 2004).

Quelques auteurs ont proposé des méthodes basées sur les modèles de chaînes de Markov cachées : Asai et al. (1993), Stultz et al. (1993); White et al. (1994), Thorne et al. (1996); Lio et al. (1998), et Crooks and Brenner (2004).

Parmi les autres approches, nous pouvons citer entre autres, les statistiques multivariées (Kanehisa, 1988), la programmation logique inductive (Muggleton et al., 1992; Palu et al., 2004) ou encore la mécanique moléculaire (Kilosanidze et al., 2004).

1.3 Les méthodes consensus

Dès 1998, des approches consistant à rechercher un consensus entre plusieurs algorithmes de prédiction ont été proposés (Biou et al., 1988). Le but est de corriger les biais intrinsèques à chaque méthode, en effectuant la prédiction par plusieurs algorithmes puis en combinant les prédictions. Ainsi, en combinant trois méthodes, Biou et al. ont montré qu'il est possible d'améliorer le Q_3 de 2,5 à 6,5 points par rapport aux prédictions des méthodes individuelles. PROF_King (Ouali and King, 2000) est une méthode consensus qui utilise des classifieurs en cascade (dont les méthodes GOR) en plus de réseaux de neurones développés par les auteurs. Le Q_3 est ainsi amélioré de plus de 10 points, atteignant 76,7%. La méthode HYDROSP (Wu et al., 2004) combine PSIPRED et une banque de petits fragments de structure connue. Ceci permet d'améliorer la prédiction de PSIPRED pour les protéines ayant une forte similarité locale avec la banque de fragments. Les travaux de Guermeur et al. (2004) utilisent les SVM multiclasses pour combiner GOR, SOPMA et SIMPA, ou les réseaux de neurones de SSPRO.

1.4 Evaluation des méthodes de prédiction de la structure secondaire

Comparer objectivement entre elles les différentes méthodes n'est pas trivial, et ceci pour plusieurs raisons.

i. **La référence utilisée**, autrement l'assignation des structures secondaires considérées comme standard de vérité, doit être identique pour toute les méthodes. L'assignation est généralement celle fournie par le programme DSSP (Kabsch and Sander, 1983). Cependant, DSSP

assigne initialement 8 conformations locales : hélice α, brin β, hélice 3-10, hélice π, coude, bend (région courbée), β-bridge isolé et coil. Les méthodes de prédiction fournissant habituellement une prédiction à trois modalités, il est nécessaire de réduire ces 8 états en 3. Il s'avère que la réduction adoptée influence les taux de performances (Rost, 2001).

ii. **Les scores utilisés** pour quantifier la prédiction doivent refléter au mieux les apports et lacunes de chaque méthode. En général, les auteurs rapportent le score Q_3. D'autres indices sont plus appropriés, comme le SOV (Zemla et al., 1999) qui tient compte du recouvrement en terme de segments de structures secondaires.

iii. **La redondance des jeux de données** utilisés pour l'entraînement et le test des méthodes peut biaiser l'évaluation des méthodes. D'une part, si le jeu d'apprentissage contient des séquences homologues, les paramètres risquent d'être biaisés vers les familles de séquences les plus représentées. D'autre part, si le jeu de test présente des homologies avec le jeu d'apprentissage, les performances risquent d'être sur-évaluées.

iv. **Les jeux de test** peuvent être intrinsèquement plus ou moins difficiles à prédire. Par exemple, la prédiction des hélices α étant plus facile, les méthodes testées sur des protéines tout α présentent de meilleurs résultats. D'une manière générale, il est délicat de comparer les méthodes évaluées sur différents jeux de données. Avec une méthode donnée, en prenant les précautions nécessaires sur la redondance, il est possible d'obtenir des performances différentes sur deux jeux de test non corrélés. Ainsi, B. Rost rapporte avoir obtenu des performances significativement différentes en testant sa méthode PROF sur deux jeux de données distincts, non corrélés entre eux, et non corrélés au jeu d'apprentissage (Rost, 2001).

v. **La date à laquelle a été effectué le test** modifie les résultats des méthodes utilisant les profils de séquences. En effet, le contenu des banques de données étant en perpetuelle augmentation, les profils s'enrichissent et les méthodes utilisant les profils voient leur performances s'améliorer par ce seul effet. Cette précaution semble parfois négligée au moment de comparer les performances d'une nouvelle méthode avec celles déjà publiées (Kim and Park, 2003).

A cet égard, un serveur a été mis en place : EVA (Koh et al., 2003). Les méthodes inscrites à EVA sont testées en continu sur les nouvelles structures déposées dans la PDB. Seules les protéines qui n'ont pas d'homologie détectable avec les protéines déjà déposées dans la PDB sont utilisées pour l'évaluation, ce qui permet d'éviter les biais de sur-apprentissage. Les structures secondaires de référence et les indices utilisés sont les mêmes pour toutes les méthodes.

En revanche le nombre de séquences analysées dépend de la date à laquelle les méthodes ont été incluses dans la comparaison, ce qui ne permet pas de comparer les performances de toutes les méthodes entre elles. Peu de détails sont donnés sur les précautions prises pour contrôler la composition en structure secondaire des séquences testées. D'autre part, cette comparaison assure artificiellement une meilleure visibilité aux méthodes participantes.

Compte-tenu de ces réserves, les résultats d'EVA montrent que, sur des jeux de données de plus de 100 structures, les meilleures méthodes, en général basées sur des réseaux de neurones, atteignent des Q_3 de 78%. Pour connaître les méthodes les plus performantes à l'heure actuelle, rendez vous sur le site d'EVA : http ://cubic.bioc.columbia.edu/eva/.

2 Au delà des structures secondaires

Nous ne nous attarderons pas sur le sujet, mais sachez qu'en complément des structures secondaires, un certain nombre d'efforts ont été entrepris pour prédire la structure des boucles protéiques (Donate et al., 1996; Oliva et al., 1997; Wojcik et al., 1999; Michalsky et al., 2003; Espadaler et al., 2004; Fernandez-Fuentes et al., 2005, 2006).

La description et la prédiction des structures secondaires ne fournit aucune indication structurale sur l'apériodique. L'utilisation d'alphabets structuraux apporte une description complète des structures. Un alphabet structural peut être vu comme un jeu de Lego ®, dont les briques permettent de reconstruire les structures 3D existantes. Les alphabets structuraux répondent à un double objectif : (i) fournir une description précise des structures, dans un but d'analyse ou de reconstruction, et (ii) être utilisés pour la prédiction de la structure locale.

Ainsi, pour échapper à la description classique à trois états, différentes méthodes génériques ont été décrites. Bystroff and Baker (1998) ont proposé les I-sites, un ensemble de conformations récurrentes au sein des structures protéiques. Ces fragments de 3 et 9 résidus de long possèdent une signature de séquence suffisante pour permettre la prédiction de conformations préférentielles de certaines régions. Pour améliorer les performances de prédiction, les I-Sites ont ensuite été inclus dans un modèle de prédiction basé sur des chaînes de Markov cachées : HMM-STR (Bystroff et al., 2000). Ce modèle de Markov permet de décrire les transitions compatibles entre les motifs séquence-structure constituant les I-sites. A partir des 28 centroïdes de leur alphabet structural (Hunter and Subramaniam, 2003b), Hunter and Subramaniam (2003a) ont défini une approche de prédiction de type bayesienne après avoir évalué la probabilité d'apparition de chaque type d'acide aminé à chaque position de leur fragments canoniques. de Brevern et al. (2000) ont mis en place une méthode de prédiction baysienne de leur 16 *Protein Blocks* (PB) (de Brevern et al., 2000) dont la performance a été améliorée (i) en élargissant la banque d'apprentissage de la relation séquence-structure et en s'aidant de méthode de prédiction de la structure secondaire (Etchebest et al., 2005), et en prennant en compte la dépendance des transitions entre les PB (de Brevern et al., 2007). En parallèle, Benros et al. (2006) ont développé une méthode de prédiction de fragments longs basée sur le modèle de la protéine hybride (de Brevern and Hazout, 2003), une méthode de classification non supervisée. Yang and Wang (2003) ont défini une méthode consensus de prédiction de fragments de 9 résidus basée sur quatre états définis par des régions clés du diagramme de Ramachandran. Sander et al. (2006) ont quant à eux mis en place une stratégie hybride de classification prenant en compte à la fois l'information de séquence et de structure pendant le partitionnement.

Deuxième partie
Prédiction de la structure tertiaire

Une vision classique dans le domaine est de découper les méthodes de prédiction de la structure des protéines en deux catégories principales : les méthodes de modélisation comparatives

et les méthodes dites *ab initio*. Les méthodes de modélisation comparative nécessitent une ou des structures matrices identifiées par homologie entre la séquence de la structure à prédire, et un ensemble de séquences de structures expérimentalement résolues. Cette catégorie regroupe la modélisation par homologie, et les méthodes dites d'enfilage ou de reconnaissance de repliement. A l'opposé, les méthodes *ab initio* permettent de prédire une structure protéique pour des pourcentages d'identité de séquence très faibles. Ce sont donc potentiellement des méthodes permettant de générer de nouveaux repliements. Au sein des méthodes *ab initio*, nous pouvons faire la distinction entre les méthodes dites *ab initio* pures, basées uniquement sur des principes physiques, et les méthodes dites *de novo* qui, quant à elles, utilisent une batterie d'informations issues de bases de données.

Devant l'enjeu de ces méthodes de prédiction, a été mise en place une procédure d'évaluation objective de leur performance. Cette évaluation a lieu tous les deux ans lors de l'expérience CASP (Moult, 1999; Moult et al., 1997, 2001, 2003, 2005, 2007). Lors de cette expérience, les organisateurs gardent secrètes des structures protéiques récemment résolues expérimentalement, et ne fournissent aux équipes participantes que la séquence en acides aminés de ces protéines. Chaque équipe doit donc appliquer sa méthode pour proposer des modèles. A la fin de l'expérience, les structures sont rendues publiques, et les différentes méthodes évaluées pour la pertinence de leurs prédictions. Cette expérience est ainsi un très bon outil d'analyse des points forts et des faiblesses d'une méthode.

3 Les méthodes de modélisation comparative

3.1 La modélisation par homologie

Historiquement, les méthodes de modélisation par homologie sont les plus anciennes, et la technique n'a que peu évolué depuis ces vingts dernières années (Tramontano and Morea, 2003). Cette méthode se base sur le principe que des séquences, étant reliées du point de vue évolutif, possèdent un repliement tridimensionnel similaire (Holm and Sander, 1996). Ainsi, une similarité de séquence suggère une similarité de repliement. Partant de ce principe, la modélisation par homologie consiste en quatre étapes principales (Sanchez and Sali, 1997; Fiser and Sali, 2003) : (i) identification, par leur séquence, de structures connues pouvant servir de matrice, (ii) alignement de la séquence à modéliser avec la structure matrice, (iii) modélisation des régions conservées, en utilisant la matrice, et des boucles et chaînes latérales absentes de la matrice, et enfin (iv) raffinement et évaluation de la qualité du modèle produit.

La performance des méthodes de modélisation par homologie dépend directement du pourcentage d'identité de séquence que partagent la cible et la matrice. Si plus de 50% de ces deux séquences sont identiques, les prédictions sont de très haute qualité, et, il a été montré que ces modèles sont aussi pertinents que des modèles RX à faible résolution (Kopp and Schwede, 2004). Entre 30 et 50% d'identité de séquence, plus de 80% des carbones alpha sont attendus comme étant à moins de 3,5 Å de leur position réelle (Kopp and Schwede, 2004), alors qu'à moins de 30% d'identité de séquence, il y a de fortes chances que le modèle contienne des erreurs importantes

(Vitkup et al., 2001; Kopp and Schwede, 2004).

Une étude récente évaluant la pertinence des différentes méthodes de modélisation par homologie peut être trouvée dans Dalton and Jackson (2007).

3.1.1 Détection d'homologues et méthodes d'alignements

Classiquement, pour des pourcentages d'identité de séquence supérieurs à 30%, l'identification d'homologues structuraux est réalisée en comparant la séquence de la cible avec l'ensemble des séquences des structures de la PDB (*Protein Data Bank*) (Berman et al., 2000). L'alignement des séquences est réalisé par des algorithmes de type programmation dynamique (Needleman and Wunsch, 1970) et ses dérivés (Smith et al., 1981). Le programme le plus couramment utilisé dans ce cas est BLAST (*Basic Local Alignment search Tool*). Pour des pourcentages d'identité de séquence plus faibles (inférieurs à 30%), des méthodes alternatives ont du être développées. Ce sont les méthodes basées sur des profils, comme par exemple, les méthodes de recherche position-spécifiques tel que PSI-BLAST (*Position Specific Iterative BLAST*) (Altschul et al., 1997) ou des chaînes de Markov cachées (HMM, *Hidden Markov Models*) (Krogh et al., 1994). En améliorant la qualité des alignements produits, ces méthodes ont rendu possible la détection d'homologues plus distants.

Les méthodes basées sur des profils de séquences (comme PSI-BLAST) commencent par réaliser une recherche d'alignements deux à deux avec une base de données. Les alignements les plus significatifs sont concervés pour construire une matrice de score spécifique de chaque position (*PSSM : Position Specific Score Matrix*). Cette matrice remplace alors la séquence de la cible pour les prochaines itérations ; le processus est itéré jusqu'a ce qu'aucun nouvel alignement significatif ne puisse être trouvé.

Les méthodes de comparaison séquence-profil peuvent être améliorées en ajoutant des informations relatives à l'évolution de la séquence protéique pour, à la fois la séquence de la cible et les séquences de la base de données. Différentes méthodes d'alignement de profils ont été proposées en ce sens (Rychlewski et al., 2000; Yona and Levitt, 2002; Sadreyev and Grishin, 2003; Edgar and Sjolander, 2003; Pei et al., 2003). Ohlson et al. (2004) ont démontré que les méthodes profil-profil sont 30% plus performantes que les méthodes séquence-profil, à la fois pour leur capacité à reconnaître des protéines de la même super-famille, mais aussi pour la qualité des alignements produits.

Depuis CASP5 (Moult et al., 2003), il est apparu que les méta-méthodes, tirant parti de différentes approches, permettent de détecter des homologues structuraux plus lointains (Kinch et al., 2003). Deux serveurs se sont ainsi illustrés lors de CASP5 dans la catégorie reconnaissance de repliement. Il s'agit de @TOME (Douguet and Labesse, 2001) et 3D-Jury (Ginalski et al., 2003a). @TOME combine les résultats des programmes PDB Blast, 3D-PSSM (Kelley et al., 2000), mGenTHREADER (Jones, 1999a), FUGUE (Shi et al., 2001), SAM-T99 (Karplus et al., 1999), et JPRED2 (Cuff et al., 1998), tandis que 3D-Jury combine les 8 méthodes ORFeus (Ginalski et al., 2003b), SAM-T02 (Karplus et al., 2003), FFAS03 (Rychlewski et al., 2000),

mGenTHREADER (Jones, 1999a), INBGU (Fischer, 2000), RAPTOR (Xu et al., 2003) FUGUE-2 (Shi et al., 2001) et 3D-PSSM (Kelley et al., 2000) pour détecter des candidats structuraux qui sont ensuite réordonnés selon un score de similarité entre les modèles basé sur l'outil MaxSub (Siew et al., 2000).

3.1.2 Génération de modèles

Etant donné un alignement entre une séquence cible et une matrice, trois types de méthodes peuvent être utilisées pour générer un modèle, dépendant de la manière dont l'information de structure est transférée à la séquence. Ces méthodes sont : l'assemblage en corps rigides, la correspondance de segments, et la satisfaction de contraintes spatiales.

Les premiers programmes de modélisation étaient basés sur des méthodes d'assemblage rigide, dans lesquelles un modèle est assemblé à partir d'un nombre restreint de corps rigides obtenus à partir du coeur des régions alignées (Blundell et al., 1987; Greer, 1990). L'assemblage consiste donc en un ajustement des corps rigides sur la matrice, puis de reconstruire les parties non conservées (*i.e.* les boucles et chaînes latérales). Les programmes les plus connus dans cette catégorie sont SWISS-MODEL (Schwede et al., 2003), NEST (Petrey et al., 2003), 3D-JIGSAW (Bates et al., 2001) et Builder (Koehl and Delarue, 1994, 1995). La principale différence entre ces méthodes réside dans la technique employée pour reconstruire les boucles et chaînes latérales. NEST par exemple utilise une approche séquentielle, appliquant un évenement évolutif à la fois, tandis que 3D-JIGSAW et Builder utilisent des méthodes de minimisation de champ moyen (Koehl and Delarue, 1996).

L'approche par correspondance de fragments utilise un sous-ensemble de fragments protéiques dérivés de l'alignement pour rechercher des fragments compatibles dans une base de données représentative de structures résolues (Jones and Thirup, 1986; Claessens et al., 1989). La base de données de recherche contient de courts fragments protéiques sélectionnés selon des critères énergétiques et/ou géométriques. SegMod/ENCAD (Levitt, 1992) dévelopée par Michael Levitt appartient à cette catégorie.

La dernière méthode utilisée en modélisation par homologie utilise des contraintes spatiales pour générer des modèles, contraintes dérivées de l'alignement de séquences toujours. Le modèle est ainsi obtenu en minimisant les violations de ces contraintes spatiales. Le programme le plus performant, et donc le plus couramment utilisé, dans cette catégorie est MODELLER (Sali and Blundell, 1993).

Il semble difficile de dire quelle méthode semble la plus pertinente étant donné que les trois programmes NEST, SegMod/ENCAD, et MODELLER, appartenant chacun à une catégorie distincte, sont aussi performants (Wallner and Elofsson, 2005).

3.2 Les méthodes dites "d'enfilage"

Les méthodes de reconnaissance de repliement reposent sur le principe que le nombre de repliements différents que les protéines peuvent adopter est bien moins conséquent que la vaste diversité des séquences générées par les projets génomes. Il a été démontré que la PDB contenait d'ores et déjà l'ensemble des repliements différents que peuvent adopter les protéines de taille allant jusqu'à 100 résidus (Kihara and Skolnick, 2003). Bien que les structures soient mieux conservées que les séquences, il semble nécessaire que la position de certains résidus spécifiques soit conservée durant le processus d'évolution pour garantir la stabilité et la fonction d'une protéine. La sensibilité des méthodes basées sur des profils a ainsi été augmentée en exploitant cette propriété. Pour ce faire, sont inclues dans les profils des informations relatives aux structures des protéines, telles que des alignements multiples de structures, l'environnement de certains résidus, la prédiction des structures secondaires ou l'accessibilité au solvant (Tang et al., 2003; Przybylski and Rost, 2004).

Les méthodes d'enfilage évaluent la pertinence d'enfiler une séquence protéique dans une structure connue issue d'une librairie de repliements. Xu and Xu (2000) ont développé un algorithme d'enfilage qui considère les paires de contacts entre les hélices α et les brins β, et permettant des *gaps* dans l'alignement au niveau des boucles. La méthode permet d'incorporer un certain nombre de contraintes à propos de la protéine cible, telle que des ponts disulfure ou des contraintes de distance. Dans une autre approche, le problème de la reconnaissance de repliement est considéré comme un problème d'optimisation globale d'une fonction d'énergie (Xu et al., 2003), résolu par programmation linéaire.

Il a été montré lors de l'expérience CASP6 que les méthodes de reconnaissance de repliements avaient fait des progrès notables par rapport aux expériences précédentes (Wang et al., 2005).

4 Les méthodes dites *ab initio*

4.1 Les méthodes *ab initio* pures

Les méthodes *ab initio* pures n'utilisent pas d'informations directement issues de bases de données. Elles tentent d'identifier, pour une protéine dans son environnement, la structure ayant l'énergie libre la plus basse, et ce en utilisant uniquement la séquence en acides aminés de cette dernière et les lois de la physique. Cette classe de prédiction de la structure des protéines peut *a priori* être utilisée pour n'importe quelle séquence protéique avec des potentiels ayant un sens physique et une représentation atomique des modèles. C'est de loin la catégorie la plus complexe, mais aussi la plus intéressante tant elle peut nous apprendre sur le repliement des protéines.

Rose et al. ont introduit une approche hiérarchique pour prédire la structure des protéines (LINUS) qui met l'accent sur le rôle des interactions stériques et de l'entropie conformationnelle (Srinivasan and Rose, 1995, 2002).
Scheraga et al. ont aussi introduit une approche hiérarchique en utilisant un champ de force simplifié pour les calculs initiaux, suivit d'une étape de raffinement avec un champ de force

tous atomes (Lee et al., 2001; Liwo et al., 1997a,b, 2001, 2002; Pillardy et al., 2001). Ce champ de force gros grain UNRES (*UNited RESidue*), réduisant la représentation des acides aminés à seulement deux sites d'interactions, permet à l'algorithme de *Conformational Space Annealing* (CSA) d'identifier des structures de plus basse énergie (Lee et al., 1997, 1998; Lee and Scheraga, 1999; Lee et al., 2000). Les travaux récents ont porté sur une amélioration de la gestion des brins β par l'algorithme (Czaplewski et al., 2004a), une analyse détaillée du rôle des ponts disulfure dans la structure (Czaplewski et al., 2004b), et l'introduction d'un algorithme de Monte-Carlo basé sur l'échange de répliques avec la minimisation du champ de force UNRES (Nanias et al., 2005).

Floudas et al. ont développé une méthode originale de prédiction *ab initio* : ASTROFOLD (Klepeis and Floudas, 2003b). Cette approche suit aussi un modèle hiérarchique : dans un premier temps, les segments en hélices sont prédits (Klepeis and Floudas, 2002), puis les régions en brins β et la topologie du feuillet correspondant sont envisagées en maximisant le nombre d'interactions hydrophobes (Klepeis and Floudas, 2003a). Un ensemble de conformères est ensuite prédit pour chaques boucles *via* des calculs d'énergie libre (Klepeis and Floudas, 2005) couplés à un échantillonnage intensif et une procédure de classification (Monnigmann and Floudas, 2005). L'ensemble de ces prédictions isolées permettent de définir des contraintes utilisées pour la prédiction de la structure tertiaire par l'intermédiaire d'une nouvelle classe hybride d'optimisation globale (Klepeis et al., 2003a,b) similaire aux protocoles de raffinement des structures RMN (Klepeis and Floudas, 1999).

Dill et al. ont proposé un autre mécanisme potentiel pour le repliement protéique : la fermeture éclair hydrophobe (ou *Zipping and Assembly*) (Dill et al., 1993). Dans ce cas, la formation de structures secondaires s'effectue en même temps que l'effondrement hydrophobe : les structures secondaires commencent à se former localement à différents endroits indépendants de la chaîne, puis ces noyaux s'assemblent pour former une structure complète (Dill et al., 2007). Ce mode de repliement permettrait d'apporter une explication au paradoxe de Levinthal (1968), *i.e.* pourquoi le phénomène de repliement est-il aussi rapide ? Ainsi, Dill et al. ont appliqué ce mécanisme de recherche dans leur méthode de prédiction *ab initio*. Ils partent d'une structure totalement dépliée, sur laquelle s'applique un algorithme basé sur des simulations de dynamique moléculaire avec échange de répliques, guidé par le champ de force AMBER96 (Kollman et al., 1997) et un modèle de solvant implicite de type born généralisé (GBSA) (Tsui and Case, 2000). Les structures à prédire sont d'abord découpées en fragments de 8 à 12 résidus, puis simulées indépendamment jusqu'à ce que se forme un nombre suffisant de contacts hydrophobes. Leur algorithme fait ensuite croître le fragment en appliquant des contraintes sur les contacts déjà formés. Ces ensembles de simulations sont répétées sur plusieurs régions de la protéine à modéliser, jusqu'à ce qu'aucun contact supplémentaire ne puisse se former. Ils ont ensuite recours à une méthode d'assemblage de fragments pour générer une structure complète. Cette méthode semble donner de bons résultats pour des protéines de petite taille : une étude récente a montré qu'ils ont pu générer des modèles éloignés en moyenne de 2,2 Å de la structure native pour huit des neuf protéines de leur jeu de test ayant des tailles comprises entre 25 et 73 acides aminés

(Ozkan et al., 2007).

Il s'avère qu'à l'heure actuelle, bien que ces méthodes donnent des résultats encourageants pour des protéines de petite taille, elles ne sont pas les plus performantes pour générer des modèles protéiques face aux méthodes *de novo*, car trop limitées par les ressources de calcul nécessaires.

4.2 Les méthodes dites *de novo*

Les méthodes dites *de novo* sont actuellement les plus performantes pour fournir des modèles à moyenne voir haute résolution (Floudas, 2007). Elles peuvent être découpées en deux catégories : (i) les méthodes d'assemblage de fragments ; (ii) les méthodes hybrides qui mêlent à la fois de l'assemblage de fragments et des simulations de repliement sur réseau.

4.2.1 Les méthodes d'assemblage de fragments

Les méthodes d'assemblage de fragments s'appuient sur le principe fondamental suivant : les interactions locales séquence-dépendantes conduisent la chaîne protéique à n'échantillonner qu'un sous-ensemble restreint de conformations, tandis que les interactions non locales préfèrent des conformères d'énergie libre compatibles avec le biais de conformères locaux (Floudas, 2007).

Dans ce type d'approche le principe est de proposer un ensemble de fragments candidats prédits à partir de la séquence en acides aminés et couvrant la totalité de cette dernière. Ces fragments sont ensuite assemblés pour former un modèle protéique complet.

La méthode la plus performante à l'heure actuelle, dans cette catégorie, est certainement Rosetta (Rohl et al., 2004; Bradley et al., 2005) et ses dérivées développées dans le groupe de David Baker. La méthode Rosetta initiale se déroule en deux étapes : (i) prédiction des fragments de 3 et 9 résidus pouvant décrire la séquence par des méthodes de comparaison de profils raffinées par différentes méthodes de prédiction des structures secondaires (dont PSIPRED (Jones, 1999b)), et (ii) génération de modèles par assemblage de fragments en utilisant un algorithme de type recuit simulé (Simons et al., 1997). Dans cette méthode d'assemblage, un grand nombre de simulations indépendantes est réalisé. Pour chaque simulation, le point de départ est une chaîne protéique totalement dépliée. A chaque pas du recuit simulé, une fenêtre sélectionnée au hasard est sujette à l'insertion d'un des meilleurs fragments de neufs résidus prédits à cette position. Cette insertion se fait en imposant les angles de torsion du fragment considéré. A chaque pas, la pertinence du modèle est évaluée par une fonction de score (Rohl et al., 2004) dont la formulation évolue au cours de la simulation (chacun des termes est associé jusqu'à la formulation complète à la fin de la simulation). A l'issue de la simulation, un raffinement du modèle est effectué par l'insertion des fragments prédits de 3 résidus avec la formulation complète de la fonction objectif. Dans son protocole standard, Rosetta utilise un modèle gros grain, dans lequel, la chaîne principale est représentée explicitement, ainsi que les carbones β, tandis que le reste de la chaîne latérale est représenté par une sphère localisée sur le centre de masse de cette

dernière (pour la glycine le carbone α est choisi comme référence). Cependant, Rosetta est aussi capable de raffiner des modèles tous atomes avec un potentiel physique. Les modèles générés sont ensuite classifiés selon leur similarité structurale, et les modèles retenus sont en général les centres des classes (Bonneau et al., 2002).

SIMFOLD développée récemment par l'équipe de Shoji Takada (Chikenji et al., 2003) est une méthode proche de Rosetta qui a donné de bons résultats lors de la $6^{\text{ème}}$ expérience CASP. Dans cette dernière, sont utilisés des fragments de 4 et 9 résidus assemblés par un algorithme très évolué de type Monte Carlo, guidé par une fonction d'énergie basée sur des considérations physiques (Fujitsuka et al., 2004). La principale originalité de la méthode réside dans son mode d'insertion des fragments qui est réversible. A la différence de Rosetta qui n'utilise que le sous-ensemble de fragments prédits pour les insérer dans le modèle, SIMFOLD insère un nouveau fragment pour faire la jonction entre deux fragments, l'ancienne conformation de la région substituée est ajoutée à l'ensemble des fragments disponibles, et peut donc être insérée à nouveau.

PROFESY est une autre approche *de novo* basée sur un assemblage de fragments mis en place au sein du groupe de Jooyoung Lee (Lee et al., 2004). Comme SIMFOLD, et à la différence des algorithmes classiques de recuit simulé, PROFESY se base sur une méthode efficace d'échantillonage de l'espace conformationnel et utilise un potentiel physique plus que statistique. La minimisation globale de la fonction d'énergie est rendue possible par l'algorithme de CSA développé au sein du groupe de Harold Scheraga (Lee et al., 1999).

FRAGFOLD, une méthode originale de David Jones, utilise une librairie de fragments constituée de super-structures secondaires (composées de deux ou trois structures secondaires consécutives) issues d'une collection de structures protéiques à haute résolution, ainsi que de petits fragments de 3 à 5 résidus de long (Jones, 1997; Jones and McGuffin, 2003). Les fragments compatibles prédits sont assignés à la séquence à prédire par une méthode d'enfilage directement inspirée de l'algorithme de GenThreader (Jones, 1999a). La structure globale est reconstruite par un algorithme génétique ou un recuit simulé, dans lesquels, la moitié des mouvements aléatoires correspondent à l'insertion de motifs de super-structures secondaires sélectionnés, et l'autre moitié à l'insertion de petits fragments. L'ensemble des modèles générés est ensuite évalué en terme de collisions stériques, de compacité et de nombre de liaisons hydrogène. Enfin, les modèles sont regroupés en classes de repliements les plus représentatifs.

4.2.2 Les méthodes hybrides

Les modèles sur réseau sont une alternative aux méthodes d'assemblage de fragments. Dans ces méthodes, l'espace conformationnel est limité à un ensemble de points que sont les points d'intersection d'une grille à trois dimensions (Skolnick and Kolinski, 1991; Hinds and Levitt, 1992). La résolution du système dépend de la taille de la maille. Ces méthodes permettent une exploration très rapide d'un grand nombre de conformations, mais souffrent d'une faible résolution et par conséquent, de la difficulté d'y implémenter des fonctions d'énergie ayant un sens

physique. Récemment, dû au succès grandissant des méthodes d'assemblage de fragments, de nouvelles méthodes hybrides sont apparues combinant les deux approches.

La méthode la plus aboutie dans cette catégorie est TASSER (Zhang et al., 2005). Mise en place par Jeffrey Skolnick et Yang Zhang, TASSER et ses méthodes dérivées ont remporté un franc succès lors de CASP6 (Zhang et al., 2005) et CASP7 (Zhou et al., 2007). La première étape de TASSER est d'identifier à la fois un repliement consensus, mais aussi un ensemble de repliements matrice distincts par la méthode d'enfilage PROSPECTOR (Skolnick et al., 2004). De par les alignements obtenus par la méthode de reconnaissance de repliement, la chaîne protéique est découpée en fragments contigus alignés d'au moins 5 résidus et en regions non alignées. La conformation des régions alignées est copiée en l'état et n'est pas changée lors de la procédure d'assemblage, alors que les régions non alignées sont repliées *ab initio* sur un réseau cubique similaire à ceux développées par Skolnick, Kolinski et al. (Kolinski et al., 2001; Kihara et al., 2001). L'ensemble des modèles générés est alors classé par la méthode itérative SPICKER (Zhang and Skolnick, 2004) identifiant les modèles selon la densité des classes générées. Deux améliorations majeures ont été récemment apportées à la méthode et testées avec succès à CASP7 : Wu et al. (2007) ont rendu la méthode itérative pour progressivement raffiner les modèles obtenus, et Zhou and Skolnick (2007) ont ajouté, à l'ensemble des fragments identifiés par reconnaissance de repliement, des super-structures secondaires repliées *ab initio* : les *chunks*.

Une autre approche hybride a été développée par Kolinski et Bujnicki, tirant partie de deux méthodes : FRankenstein's Monster (FRM) (Kosinski et al., 2003) et CABS (*Cα-β and Side group*) (Kolinski, 2004). Dans un premier temps, des modèles hybrides sont générés avec la méthode FRM de recombinaison de matrices, puis classés selon leur score Verify3D (Bowie et al., 1991; Luthy et al., 1992) pour identifier les fragments correctement repliés. Ces fragments ne sont pas utilisés directement, mais sont utilisés comme source de contraintes spatiales permettant de guider les simulations de Monte-Carlo avec échange de répliques du modèle CABS. L'ensemble des modèles générés est ensuite classé par la méthode HCPM (Gront and Kolinski, 2005) (*Hierarchical Clustering of Protein Models*) pour identifier les modèles finaux. Cette méthode performante a donné de bons résultats lors de la sixième édition de CASP (Kolinski and Bujnicki, 2005).

Si l'on fait un bilan des performances de l'ensemble des méthodes de prédiction de la structure des protéines, il apparait que les méthodes de modélisation comparative sont très efficaces depuis quelques années et ne semblent de ce fait que très peu évoluer. Les méthodes d'assemblage de fragments continuent de s'améliorer, mais il apparaît que des méthodes hybrides, telle TASSER, combinant à la fois de l'enfilage et une méthode d'assemblage de fragments, puisse rapidement progresser pour donner des résultats très pertinent dans un avenir proche. @TOME et 3D-Jury ont ouvert la porte aux méta-méthodes de prédiction de la structure des protéines. Un méta-serveur, combinant les approches d'assemblage de fragments les plus performantes à l'heure actuelle, serait bienvenue pour parfaire notre connaissance du repliement protéique, et fournir aux biologistes un outil de confiance.

Troisième partie

Les outils et serveurs conseillés

5 Les serveurs de prédiction

Voici une liste non exhaustive des serveurs les plus réputés à l'heure actuelle. Notez que certains serveurs peuvent être utilisés à des fins différents, les catégories proposées ici ne sont pas fixes.

5.1 Prédiction des structures secondaires

- ▷ **PSIPRED** :
  http ://bioinf.cs.ucl.ac.uk/psipred/
- ▷ **SSPRO** :
  http ://contact.ics.uci.edu/sspro4.html
  Considérez aussi :
  http ://scratch.proteomics.ics.uci.edu/
- ▷ **SAM-T99sec** :
  http ://www.soe.ucsc.edu/research/compbio/HMM-apps/T99-query.html
- ▷ **PROFsec** et **PHDsec** : la soumission de séquences à ce serveur nécessite la création d'un compte gratuit. Ce serveur est une véritable boîte à outils : recherche de patterns, définition de domaines, prédiction de l'oxydation des cystéines, de la compacité de la structure, etc.
  http ://www.predictprotein.org/
- ▷ **PROF_King** :
  http ://www.aber.ac.uk/ phiwww/prof/

5.2 Recherche d'informations

- ▷ **Ponts disulfure** (Connectivité des cystéines)
  - **DIANA** :
    http ://clavius.bc.edu/ clotelab/DiANNA/
  - **GDAP** :
    http ://www.doe-mbi.ucla.edu/ boconnor/GDAP/new_query.php
  - **CysState** :
    http ://bioserv.rpbs.jussieu.fr/
- ▷ **Segments tranmembranaires**
  - **Split4** :
    http ://split.pmfst.hr/split/4/
  - **RMPred** :
    http ://www.ch.embnet.org/software/TMPRED_form.html
- ▷ **Homologie**
  - **3DJury** :
    http ://meta.bioinfo.pl/submit_wizard.pl

- **HHPred** :
  http ://toolkit.tuebingen.mpg.de/hhpred

▷ **Fonction**

- **PFP** :
  http ://dragon.bio.purdue.edu/pfp/
- **ProtFun** :
  http ://www.cbs.dtu.dk/services/ProtFun/
- Voir les méthodes de recherche d'homologues

5.3 Prédiction de la structure tertiaire

5.3.1 Modélisation par homologie

▷ **MODELLER** : la soumission d'un job nécessite la clé de MODELLER.
http ://toolkit.tuebingen.mpg.de/modeller

▷ **SWISS-MODEL** : une véritable boîte à modéliser allant de la recherche de *templates* à la minimisation des modèles avec `Gromos96`.
http ://swissmodel.expasy.org/

▷ **Geno3D** :
http ://geno3d-pbil.ibcp.fr/

5.3.2 Enfilage (*threading*)

▷ **mGenTHREADER** : combine GenTHREADER et PSIPRED.
http ://bioinf.cs.ucl.ac.uk/psipred/

▷ **I-TASSER** : nécéssite la création d'un compte.
http ://zhang.bioinformatics.ku.edu/I-TASSER/

▷ **TASSER-Lite** :
http ://cssb.biology.gatech.edu/skolnick/webservice/tasserlite/index.html

▷ **3DJury** inclu 3D-PSSM, FUGUE, etc (Voir plus haut).

5.3.3 Ab initio

▷ **Robetta** : prévoir de le lancer à l'avance, la queue est longue !
http ://bioinf.cs.ucl.ac.uk/psipred/.

▷ **HMMStr/Rosetta** :
http ://www.bioinfo.rpi.edu/ bystrc/hmmstr/server.php

6 La boîte à outils

6.1 Evaluation des modèles

▷ **Eval123D** :
http ://bioserv.cbs.cnrs.fr/HTML_BIO/valid.html

▷ **Procheck**

- ▷ **Verify3D** :
  http ://nihserver.mbi.ucla.edu/Verify_3D/

6.2 Superposition de structures

- ▷ **iSuperpose** :
  http ://bioserv.rpbs.jussieu.fr/cgi-bin/iSuperpose/
- ▷ **ProFit** : la documentation est disponible ici :
  http ://www.bioinf.org.uk/software/profit/index.html

6.3 Analyse des modèles

- ▷ **ASA** : surface accessible au solvant
  http ://bioserv.rpbs.jussieu.fr/
- ▷ **PCE** : potentiel electrostatique
  http ://bioserv.rpbs.jussieu.fr/cgi-bin/PCE-Pot
- ▷ **el nemo** : dynamique des structures par analyse en mode normaux
  http ://www.igs.cnrs-mrs.fr/elnemo/

6.4 Reconstruction de modèles incomplets

- ▷ **SCWRL** : reconstruction des chaînes latérales
  http ://bioserv.rpbs.jussieu.fr/
- ▷ **SABBAC** : reconstruction de modèles protéiques complets à partir de leur trace en carbones alpha
  http ://bioserv.rpbs.jussieu.fr/cgi-bin/SABBAC

a propos

En dehors de mes travaux de thèse, cette introduction à la structure des protéines a été facilitée par les travaux initiateurs d'Adrien Melquiond, et à l'aimable authorisation de Juliette Martin de pouvoir reproduire certaines parties de son manuscrit de thèse disponible sur son site personnel http ://juliettemartin.fr.

Références

S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res*, 25 (17) :3389–402, September 1997.

C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96) :223–30, July 1973.

K Asai, S Hayamizu, and K Handa. Prediction of protein secondary structure by the hidden markov model. *Comput Appl Biosci*, 9(2) :141–6, April 1993. ISSN 0266-7061.

P Baldi, S Brunak, P Frasconi, G Soda, and G Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11) :937–46, November 1999.

R L Baldwin. Intermediates in protein folding reactions and the mechanism of protein folding. *Annu Rev Biochem*, 44 :453–75, 1975.

R L Baldwin and G D Rose. Is protein folding hierarchic ? i. local structure and peptide folding. *Trends Biochem Sci*, 24(1) :26–33, January 1999.

P A Bates, L A Kelley, R M MacCallum, and M J Sternberg. Enhancement of protein modeling by human intervention in applying the automatic programs 3d-jigsaw and 3d-pssm. *Proteins*, Suppl 5 :39–46, 2001.

C Benros, A G de Brevern, C Etchebest, and S Hazout. Assessing a novel approach for predicting local 3d protein structures from sequence. *Proteins*, 62(4) :865–80, March 2006.

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res*, 28(1) :235–42, January 2000.

V Biou, J F Gibrat, J M Levin, B Robson, and J Garnier. Secondary structure prediction : combination of three different methods. *Protein Eng*, 2(3) :185–91, September 1988. ISSN 0269-2139.

T L Blundell, B L Sibanda, M J Sternberg, and J M Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326(6111) :347–52, March 1987.

R Bonneau, C E M Strauss, C A Rohl, D Chivian, P Bradley, L Malmstrom, T Robertson, and D Baker. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol*, 322(1) :65–78, September 2002.

J U Bowie, R Luthy, and D Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016) : 164–70, July 1991.

P Bradley, L Malmstrom, B Qian, J Schonbrun, D Chivian, D E Kim, J Meiler, K M S Misura, and D Baker. Free modeling with rosetta in casp6. *Proteins*, 61 Suppl 7 :128–34, 2005.

C Bystroff and D Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, 281(3) :565–77, August 1998.

C Bystroff, V Thorsson, and D Baker. Hmmstr : a hidden markov model for local sequence-structure correlations in proteins. *J Mol Biol*, 301 (1) :173–90, August 2000.

G Chikenji, Y Fujitsuka, and S Takada. A reversible fragment assembly method for de novo protein structure prediction. *J Chem Phys*, 119 : 6895–6903, 2003.

P Y Chou and G D Fasman. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2) :211–22, January 1974. ISSN 0006-2960.

M Claessens, E Van Cutsem, I Lasters, and S Wodak. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng*, 2(5) :335–45, January 1989.

G E Crooks and S E Brenner. Protein secondary structure : entropy, correlations and prediction. *Bioinformatics*, 20(10) :1603–11, July 2004.

J A Cuff, M E Clamp, A S Siddiqui, M Finlay, and G J Barton. Jpred : a consensus secondary structure prediction server. *Bioinformatics*, 14(10) : 892–3, 1998.

C Czaplewski, A Liwo, J Pillardy, S Oldziej, and H A Scheraga. Improved conformational space annealing method to treat beta-structure with the unres force-field and to enhance scalability of parallel implementation. *Polymer*, 45 :677–686, 2004a.

C Czaplewski, S Oldziej, A Liwo, and H A Scheraga. Prediction of the structures of proteins with the unres force field, including dynamic formation and breaking of disulfide bonds. *Protein Eng Des Sel*, 17(1) :29–36, January 2004b.

J A R Dalton and R M Jackson. An evaluation of automated homology modelling methods at low target template sequence similarity. *Bioinformatics*, 23(15) :1901–8, August 2007.

A G de Brevern and S Hazout. 'hybrid protein model' for optimally defining 3d protein structure fragments. *Bioinformatics*, 19(3) :345–53, February 2003.

A G de Brevern, C Etchebest, and S Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41(3) :271–87, November 2000.

A G de Brevern, C Etchebest, C Benros, and S Hazout. "pinning strategy" : a novel approach for predicting the backbone structure in terms of protein blocks from sequence. *J Biosci*, 32(1) : 51–70, January 2007.

K A Dill, K M Fiebig, and H S Chan. Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci U S A*, 90(5) :1942–6, March 1993.

K A Dill, S B Ozkan, T R Weikl, J D Chodera, and V A Voelz. The protein folding problem : when will it be solved ? *Curr Opin Struct Biol*, 17(3) : 342–6, June 2007.

L E Donate, S D Rufino, L H Canard, and T L Blundell. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures : a database for modeling and prediction. *Protein Sci*, 5(12) :2600–16, December 1996.

D Douguet and G Labesse. Easier threading through web-based comparisons and cross-validations. *Bioinformatics*, 17(8) :752–3, August 2001.

H J Dyson and P E Wright. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, 6(3) :197–208, March 2005.

R C Edgar and K Sjolander. Satchmo : sequence alignment and tree construction using hidden markov models. *Bioinformatics*, 19(11) :1404–11, July 2003.

R J Ellis. Moledular chaperones ten years. introduction. *Semin Cell Dev Biol*, 11(1) :1–5, February 2000.

R J Ellis. Molecular chaperones : assisting assembly in addition to folding. *Trends Biochem Sci*, 31(7) :395–401, July 2006.

J Espadaler, N Fernandez-Fuentes, A Hermoso, E Querol, F X Aviles, M J E Sternberg, and B Oliva. Archdb : automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res*, 32(Database issue) :D185–8, January 2004.

C Etchebest, C Benros, S Hazout, and A G de Brevern. A structural alphabet for local protein structures : improved prediction methods. *Proteins*, 59(4) :810–27, June 2005.

N Fernandez-Fuentes, E Querol, F X Aviles, M J E Sternberg, and B Oliva. Prediction of the conformation and geometry of loops in globular proteins : testing archdb, a structural classification of loops. *Proteins*, 60(4) :746–57, September 2005.

N Fernandez-Fuentes, J Zhai, and A Fiser. Archpred : a template based loop structure prediction server. *Nucleic Acids Res*, 34(Web Server issue) : W173–6, July 2006.

A R Fersht and V Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108(4) :573–82, February 2002.

A Figureau, M A Soto, and J Tohá. A pentapeptide-based method for protein secondary structure prediction. *Protein Eng*, 16(2) :103–7, February 2003. ISSN 0269-2139.

D Fischer. Hybrid fold recognition : combining sequence derived properties with evolutionary information. *Pac Symp Biocomput*, pages 119–30, 2000.

A Fiser and A Sali. *Comparative protein structure modeling.*, pages 167–206. Marcel Dekker, Inc., 2003.

C A Floudas. Computational methods in protein structure prediction. *Biotechnol Bioeng*, 97(2) : 207–13, June 2007.

Y Fujitsuka, S Takada, Z A Luthey-Schulten, and P G Wolynes. Optimizing physical energy functions for protein folding. *Proteins*, 54(1) :88–103, January 2004.

J Garnier, D J Osguthorpe, and B Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*, 120(1) :97–120, March 1978.

J Garnier, J F Gibrat, and B Robson. Gor method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*, 266 : 540–53, 1996.

C Geourjon and G Deleage. Sopma : significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci*, 11(6) :681–4, December 1995. ISSN 0266-7061.

J F Gibrat, J Garnier, and B Robson. Further developments of protein secondary structure prediction using information theory. new parameters and consideration of residue pairs. *J Mol Biol*, 198(3) :425–43, December 1987.

K Ginalski, A Elofsson, D Fischer, and L Rychlewski. 3d-jury : a simple approach to improve protein structure predictions. *Bioinformatics*, 19 (8) :1015–8, May 2003a.

K Ginalski, J Pas, L S Wyrwicz, M von Grotthuss, J M Bujnicki, and L Rychlewski. Orfeus : Detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res*, 31(13) :3804–7, July 2003b.

J Greer. Comparative modeling methods : application to the family of the mammalian serine proteases. *Proteins*, 7(4) :317–34, 1990.

D Gront and A Kolinski. Hcpm program for hierarchical clustering of protein models. *Bioinformatics*, 21(14) :3179–80, July 2005.

Y Guermeur, G Pollastri, A Elisseeff, H Zelus, D Paugam-Moisy, and P Baldi. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing*, pages 305–327, 2004.

N Haspel, C J Tsai, H Wolfson, and R Nussinov. Hierarchical protein folding pathways : a computational study of protein fragments. *Proteins*, 51 (2) :203–15, May 2003.

D A Hinds and M Levitt. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci U S A*, 89(7) :2536–40, April 1992.

L Holm and C Sander. Mapping the protein universe. *Science*, 273(5275) :595–603, August 1996.

H J Hu, Y Pan, R Harrison, and P C Tai. Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier. *IEEE Trans. Nanobioscience*, 3(4) :265–271, 2004.

S Hua and Z Sun. A novel method of protein secondary structure prediction with high segment overlap measure : support vector machine approach. *J Mol Biol*, 308(2) :397–407, April 2001. ISSN 0022-2836.

C G Hunter and S Subramaniam. Protein local structure prediction from sequence. *Proteins*, 50 (4) :572–9, March 2003a.

C G Hunter and S Subramaniam. Protein fragment clustering and canonical local shapes. *Proteins*, 50(4) :580–8, March 2003b.

F Jiang. Prediction of protein secondary structure with a reliability score estimated by local sequence clustering. *Protein Eng*, 16(9) :651–657, 2003.

D T Jones. Genthreader : an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, 287(4) :797–815, April 1999a.

D T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2) :195–202, September 1999b.

D T Jones. Successful ab initio prediction of the tertiary structure of nk-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*, Suppl 1 :185–91, 1997.

David T Jones and Liam J McGuffin. Assembling novel protein folds from super-secondary structural fragments. *Proteins*, 53 Suppl 6 :480–5, 2003.

T A Jones and S Thirup. Using known substructures in protein model building and crystallography. *EMBO J*, 5(4) :819–22, April 1986.

W Kabsch and C Sander. Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12) :2577–637, December 1983. ISSN 0006-3525.

M Kanehisa. A multivariate analysis method for discriminating protein secondary structural segments. *Protein Eng*, 2(2) :87–92, July 1988. ISSN 0269-2139.

K Karplus, C Barrett, M Cline, M Diekhans, L Grate, and R Hughey. Predicting protein structure using only sequence information. *Proteins*, Suppl 3 :121–5, 1999.

K Karplus, R Karchin, J Draper, J Casper, Y Mandel-Gutfreund, M Diekhans, and R Hughey. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, 53 Suppl 6 :491–6, 2003.

M Karplus and D L Weaver. Protein folding dynamics : the diffusion-collision model and experimental data. *Protein Sci*, 3(4) :650–68, April 1994.

L A Kelley, R M MacCallum, and M J Sternberg. Enhanced genome annotation using structural profiles in the program 3d-pssm. *J Mol Biol*, 299 (2) :499–520, June 2000.

D Kihara and J Skolnick. The pdb is a covering set of small protein structures. *J Mol Biol*, 334(4) : 793–802, December 2003.

D Kihara, H Lu, A Kolinski, and J Skolnick. Touchstone : an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci U S A*, 98 (18) :10125–30, August 2001.

G T Kilosanidze, A S Kutsenko, N G Esipova, and V G Tumanyan. Analysis of forces that determine helix formation in alpha-proteins. *Protein Sci*, 13 (2) :351–7, February 2004. ISSN 0961-8368.

H Kim and H Park. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng*, 16(8) :553–60, August 2003. ISSN 0269-2139.

S Kim. Protein beta-turn prediction using nearest-neightbor method. *Bioinformatics*, 20(1) :40–4, 2004.

L N Kinch, J O Wrabl, S S Krishna, I Majumdar, R I Sadreyev, Y Qi, J Pei, H Cheng, and N V Grishin. Casp5 assessment of fold recognition target predictions. *Proteins*, 53 Suppl 6 :395–409, 2003.

J L Klepeis and C A Floudas. Analysis and prediction of loop segments in protein structures. *Comp Chem Eng*, 29 :423–436, 2005.

J L Klepeis and C A Floudas. Free energy calculations for peptides via deterministic global optimization. *J Chem Phys*, 110 :7491–7512, 1999.

J L Klepeis and C A Floudas. Ab initio prediction of helical segments in polypeptides. *J Comput Chem*, 23(2) :245–66, January 2002.

J L Klepeis and C A Floudas. Prediction of beta-sheet topology and disulfide bridges in polypeptides. *J Comput Chem*, 24(2) :191–208, January 2003a.

J L Klepeis and C A Floudas. Astro-fold : a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys J*, 85(4) :2119–46, October 2003b.

J L Klepeis, M J Pieja, and C A Floudas. A new class of hybrid global optimization algorithms for peptide structure prediction : integrated hybrids. *Comp Phys Comm*, 151 :121–140, 2003a.

J L Klepeis, M J Pieja, and C A Floudas. Hybrid global optimization algorithms for protein structure prediction : alternating hybrids. *Biophys J*, 84(2 Pt 1) :869–82, February 2003b.

P Koehl and M Delarue. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat Struct Biol*, 2(2) :163–70, February 1995.

P Koehl and M Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol*, 239(2) :249–75, June 1994.

P Koehl and M Delarue. Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol*, 6(2) :222–6, April 1996.

Ingrid Y Y Koh, Volker A Eyrich, Marc A Marti-Renom, Dariusz Przybylski, Mallur S Madhusudhan, Narayanan Eswar, Osvaldo Graña, Florencio Pazos, Alfonso Valencia, Andrej Sali, and Burkhard Rost. Eva : Evaluation of protein structure prediction servers. *Nucleic Acids Res*, 31 (13) :3311–5, July 2003. ISSN 1362-4962.

A Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*, 51(2) :349–71, 2004.

A Kolinski and J M Bujnicki. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, 61 Suppl 7 :84–90, 2005.

A Kolinski, M R Betancourt, D Kihara, P Rotkiewicz, and J Skolnick. Generalized comparative modeling (genecomp) : a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins*, 44(2) :133–49, August 2001.

P A Kollman, R Dixon, W Cornell, T Fox, C Chipot, and A Pohorille. *The Development/Application of a 'Minimalist' Organic/Biochemical Molecular Mechanic Force Field using a Combination of ab Initio Calculations and Experimental Data*, volume 3, pages 83–96. Escom, The Netherlands, 1997.

J Kopp and T Schwede. Automated protein structure homology modeling : a progress report. *Pharmacogenomics*, 5(4) :405–16, June 2004.

J Kosinski, I A Cymerman, M Feder, M A Kurowski, J M Sasin, and J M Bujnicki. A "frankenstein's monster" approach to comparative modeling : merging the finest fragments of fold-recognition models and iterative model refinement aided by 3d structure evaluation. *Proteins*, 53 Suppl 6 :369–79, 2003.

A Krogh, M Brown, I S Mian, K Sjolander, and D Haussler. Hidden markov models in computational biology. applications to protein modeling. *J Mol Biol*, 235(5) :1501–31, February 1994.

J Lee and H A Scheraga. Conformational space annealing by parallel computations : Extensive

conformational search of met-enkephalin and of the 20-residue membrane-bound portion of melittin. *Intl J of Quantum Chem*, 75 :255–265, 1999.

J Lee, H A Scheraga, and S Rackovsky. New optimization method for conformational energy calculations on polypeptides : Conformational space annealing. *J Comput Chem*, 18 :1222–1232, 1997.

J Lee, H A Scheraga, and S Rackovsky. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers*, 46(2) :103–16, August 1998.

J Lee, A Liwo, D R Ripoll, J Pillardy, and H A Scheraga. Calculation of protein conformation by global optimization of a potential energy function. *Proteins*, Suppl 3 :204–8, 1999.

J Lee, J Pillardy, C Czaplewski, Y A Arnautova, D R Ripoll, A Liwo, K D Gibson, R J Wawak, and H A Scheraga. Efficient parallel algorithms in global optimization of potential energy functions. *Comp Phys Comm*, 128 :399–411, 2000.

J Lee, D R Ripoll, C Czaplewski, J Pillardy, W J Wedemeyer, and H A Scheraga. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. *J Phys Chem B*, 105 :2323–2347, 2001.

J Lee, S Y Kim, K Joo, I Kim, and J Lee. Prediction of protein tertiary structure using profesy, a novel method based on fragment assembly and conformational space annealing. *Proteins*, 56(4) : 704–14, September 2004.

A M Lesk and G D Rose. Folding units in globular proteins. *Proc Natl Acad Sci U S A*, 78(7) : 4304–8, July 1981.

J M Levin. Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng*, 10(7) :771–6, July 1997. ISSN 0269-2139.

J M Levin, B Robson, and J Garnier. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett*, 205(2) : 303–8, September 1986. ISSN 0014-5793.

C Levinthal. Are there pathways for protein folding ? *Journal de Chimie Physique et de Physico-Chimie Biologique*, 65 :44–45, 1968.

M Levitt. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol*, 226(2) :507–33, July 1992.

M Levitt and A Warshel. Computer simulation of protein folding. *Nature*, 253(5494) :694–8, February 1975.

P Lio, N Goldman, J L Thorne, and D T Jones. Passml : combining evolutionary inference and protein secondary structure prediction. *Bioinformatics*, 14(8) :726–733, 1998.

A Liwo, S Oldziej, M R Pincus, NewAuthor3, S Rackovsky, and H A Scheraga. A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comput Chem*, 18 :849–873, 1997a.

A Liwo, S Oldziej, M R Pincus, NewAuthor3, S Rackovsky, and H A Scheraga. A united-residue force field for off-lattice protein-structure simulations. ii. parameterization of short-range interactions and determination of weights of energy terms by z-score optimization. *J Comput Chem*, 18 :874–887, 1997b.

A Liwo, C Czaplewski, J Pillardy, and H A Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J Chem Phys*, 115 :2323–2347, 2001.

A Liwo, P Arlukowicz, C Czaplewski, S Oldziej, J Pillardy, and H A Scheraga. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape : application to the unres force field. *Proc Natl Acad Sci U S A*, 99(4) :1937–42, February 2002.

R Luthy, J U Bowie, and D Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364) :83–5, March 1992.

E Michalsky, A Goede, and R Preissner. Loops in proteins (lip)–a comprehensive loop database for

homology modelling. *Protein Eng*, 16(12) :979–85, December 2003.

M Monnigmann and C A Floudas. Protein loop structure prediction with flexible stem geometries. *Proteins*, 61(4) :748–62, December 2005.

J Moult. Predicting protein three-dimensional structure. *Curr Opin Biotechnol*, 10(6) :583–8, December 1999.

J Moult, T Hubbard, S H Bryant, K Fidelis, and J T Pedersen. Critical assessment of methods of protein structure prediction (casp) : round ii. *Proteins*, Suppl 1 :2–6, 1997.

J Moult, K Fidelis, A Zemla, and T Hubbard. Critical assessment of methods of protein structure prediction (casp) : round iv. *Proteins*, Suppl 5 : 2–7, 2001.

J Moult, K Fidelis, A Zemla, and T Hubbard. Critical assessment of methods of protein structure prediction (casp)-round v. *Proteins*, 53 Suppl 6 : 334–9, 2003.

J Moult, K Fidelis, B Rost, T Hubbard, and A Tramontano. Critical assessment of methods of protein structure prediction (casp)–round 6. *Proteins*, 61 Suppl 7 :3–7, 2005.

J Moult, K Fidelis, A Kryshtafovych, B Rost, T Hubbard, and A Tramontano. Critical assessment of methods of protein structure prediction-round vii. *Proteins*, October 2007.

S Muggleton, R D King, and M J Sternberg. Protein secondary structure prediction using logic-based machine learning. *Protein Eng*, 5(7) :647–57, October 1992. ISSN 0269-2139.

M Nanias, M Chinchio, S Oldziej, C Czaplewski, and H A Scheraga. Protein structure prediction with the unres force-field using replica-exchange monte carlo-with-minimization ; comparison with mcm, csa, and cfmc. *J Comput Chem*, 26(14) : 1472–86, November 2005.

S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3) :443–53, March 1970.

T Ohlson, B Wallner, and A Elofsson. Profile-profile methods provide improved fold-recognition : a study of different profile-profile alignment methods. *Proteins*, 57(1) :188–97, October 2004.

B Oliva, P A Bates, E Querol, F X Aviles, and M J Sternberg. An automated classification of the structure of protein loops. *J Mol Biol*, 266 (4) :814–30, March 1997.

M Ouali and R D King. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci*, 9(6) :1162–76, June 2000. ISSN 0961-8368.

S B Ozkan, G A Wu, J D Chodera, and K A Dill. Protein folding by zipping and assembly. *Proc Natl Acad Sci U S A*, 104(29) :11987–92, July 2007.

A Dal Palu, A Dovier, and F Fogolari. Constraint logic programming approach to protein structure prediction. *BMC Bioinformatics*, 5 :186, November 2004. ISSN 1471-2105.

J Pei, R Sadreyev, and N V Grishin. Pcma : fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, 19(3) :427–8, February 2003.

D Petrey, Z Xiang, C L Tang, L Xie, M Gimpelev, T Mitros, C S Soto, S G Fischman, A Kernytsky, A Schlessinger, I Y Y Koh, E Alexov, and B Honig. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, 53 Suppl 6 :430–5, 2003.

J Pillardy, C Czaplewski, A Liwo, W J Wedemeyer, J Lee, D R Ripoll, P Arlukowicz, S Oldziej, Y A Arnautova, and H A Scheraga. Development of physics-based energy functions that predict medium-resolution structures for proteins of the alpha, beta and alpha/beta structural classes. *J Phys Chem B*, 105 :7299–7311, 2001.

G Pollastri, D Przybylski, B Rost, and P Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47(2) : 228–35, May 2002.

D Przybylski and B Rost. Improving fold recognition without folds. *J Mol Biol*, 341(1) :255–69, July 2004.

N Qian and T J Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol*, 202(4) :865–84, August 1988. ISSN 0022-2836.

S K Riis and A Krogh. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J Comput Biol*, 3(1) :163–83, 1996. ISSN 1066-5277.

B Robson and E Suzuki. Conformational properties of amino acid residues in globular proteins. *J Mol Biol*, 107(3) :327–56, November 1976. ISSN 0022-2836.

C A Rohl, C E M Strauss, K M S Misura, and D Baker. Protein structure prediction using rosetta. *Methods Enzymol*, 383 :66–93, 2004.

B Rost. Review : protein secondary structure prediction continues to rise. *J Struct Biol*, 134(2-3) : 204–18, 2001. ISSN 1047-8477.

B Rost. Prediction in 1d : secondary structure, membrane helices, and accessibility. *Methods Biochem Anal*, 44 :559–87, 2003. ISSN 0076-6941.

B Rost and C Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19(1) :55–72, May 1994.

L Rychlewski, L Jaroszewski, W Li, and A Godzik. Comparison of sequence profiles. strategies for structural predictions using sequence information. *Protein Sci*, 9(2) :232–41, February 2000.

R Sadreyev and N Grishin. Compass : a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, 326(1) :317–36, February 2003.

A A Salamov and V V Solovyev. Protein secondary structure prediction using local alignments. *J Mol Biol*, 268(1) :31–6, April 1997. ISSN 0022-2836.

A Sali and T L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3) :779–815, December 1993.

R Sanchez and A Sali. Advances in comparative protein-structure modelling. *Curr Opin Struct Biol*, 7(2) :206–14, April 1997.

C Sander and R Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1) : 56–68, 1991. ISSN 0887-3585.

O Sander, I Sommer, and T Lengauer. Local protein structure prediction using discriminative models. *BMC Bioinformatics*, 7 :14, 2006.

G E Schulz. Structural rules for globular proteins. *Angew Chem Int Ed Engl*, 16 :23–32, 1977.

T Schwede, J Kopp, N Guex, and M C Peitsch. Swiss-model : An automated protein homology-modeling server. *Nucleic Acids Res*, 31(13) : 3381–5, July 2003.

J Shi, T L Blundell, and K Mizuguchi. Fugue : sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, 310(1) :243–57, June 2001.

N Siew, A Elofsson, L Rychlewski, and D Fischer. Maxsub : an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9) :776–85, September 2000.

K T Simons, C Kooperberg, E Huang, and D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268(1) :209–25, April 1997.

J Skolnick and A Kolinski. Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol*, 221(2) :499–531, September 1991.

J Skolnick, D Kihara, and Y Zhang. Development and large scale benchmark testing of the prospector 3 threading algorithm. *Proteins*, 56(3) : 502–18, August 2004.

T F Smith, M S Waterman, and W M Fitch. Comparative biosequence metrics. *J Mol Evol*, 18(1) : 38–46, 1981.

R Srinivasan and G D Rose. Ab initio prediction of protein structure using linus. *Proteins*, 47(4) : 489–95, June 2002.

R Srinivasan and G D Rose. Linus : a hierarchic procedure to predict the fold of a protein. *Proteins*, 22(2) :81–99, June 1995.

C M Stultz, J V White, and T F Smith. Structural analysis based on state-space modeling. *Protein Sci*, 2(3) :305–14, March 1993. ISSN 0961-8368.

C L Tang, L Xie, I Y Y Koh, S Posy, E Alexov, and B Honig. On the role of structural information in remote homology detection and sequence alignment : new methods using hybrid sequence profiles. *J Mol Biol*, 334(5) :1043–62, December 2003.

J L Thorne, N Goldman, and D T Jones. Combining protein evolution and secondary structure. *Mol Biol Evol*, 13(5) :666–73, May 1996.

A Tramontano and V Morea. Assessment of homology-based predictions in casp5. *Proteins*, 53 Suppl 6 :352–68, 2003.

C J Tsai, P P de Laureto, A Fontana, and R Nussinov. Comparison of protein fragments identified by limited proteolysis and by computational cutting of proteins. *Protein Sci*, 11(7) :1753–70, July 2002.

V Tsui and D A Case. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers*, 56(4) : 275–91, 2000.

D Vitkup, E Melamud, J Moult, and C Sander. Completeness in structural genomics. *Nat Struct Biol*, 8(6) :559–66, June 2001.

B Wallner and A Elofsson. All are not equal : a benchmark of different homology modeling programs. *Protein Sci*, 14(5) :1315–27, May 2005.

G Wang, Y Jin, and R L Dunbrack. Assessment of fold recognition predictions in casp6. *Proteins*, 61 Suppl 7 :46–66, 2005.

J J Ward, L J McGuffin, B F Buxton, and D T Jones. Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13) : 1650–5, September 2003.

D B Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, 70(3) :697–701, March 1973.

J V White, C M Stultz, and T F Smith. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Math Biosci*, 119(1) :35–75, January 1994. ISSN 0025-5564.

J Wojcik, J P Mornon, and J Chomilier. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol*, 289(5) :1469–90, June 1999.

K P Wu, H N Lin, J M Chang, T Y Sung, and W I Hsu. Hyprosp : a hybrid protein secondary structure prediction algorithm-a knowledge-based approach. *Nucleic Acids Res*, 32(17) :5059–5065, 2004.

S Wu, J Skolnick, and Y Zhang. Ab initio modeling of small proteins by iterative tasser simulations. *BMC Biol*, 5 :17, 2007.

J Xu, M Li, D Kim, and Y Xu. Raptor : optimal protein threading by linear programming. *J Bioinform Comput Biol*, 1(1) :95–117, April 2003.

Y Xu and D Xu. Protein threading using prospect : design and evaluation. *Proteins*, 40(3) :343–54, August 2000.

A S Yang and L Y Wang. Local structure prediction with local structure-based sequence profiles. *Bioinformatics*, 19(10) :1267–74, July 2003.

G Yona and M Levitt. Within the twilight zone : a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, 315(5) :1257–75, February 2002.

A Zemla, C Venclovas, K Fidelis, and B Rost. A modified definition of sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, 34(2) :220–3, February 1999. ISSN 0887-3585.

Y Zhang and J Skolnick. Spicker : a clustering approach to identify near-native protein folds. *J Comput Chem*, 25(6) :865–71, April 2004.

Y Zhang, A K Arakaki, and J Skolnick. Tasser : an automated method for the prediction of protein tertiary structures in casp6. *Proteins*, 61 Suppl 7 :91–8, 2005.

H Zhou and J Skolnick. Ab initio protein structure prediction using chunk-tasser. *Biophys J*, 93(5) : 1510–8, September 2007.

H Zhou, S B Pandit, S Y Lee, J Borreguero, H Chen, L Wroblewska, and J Skolnick. Analysis of tasser-based casp7 protein structure prediction results. *Proteins*, August 2007.

B H Zimm and J K Bragg. Theory of the phase transition between helix and random coil in polypeptide chains. *J Chem Phys*, 31 :526–531, 1959.

K Zimmermann. When awaiting 'bio' champollion : dynamic programming regularization of the protein secondary structure predictions. *Protein Eng*, 7(10) :1197–202, October 1994. ISSN 0269-2139.