Modélisation moléculaire par homologie des protéines : ses applications en Biologie et en Bioinformatique

Jean-Luc Pellequer

Mémoire présenté en vue du diplôme d'habilitation
à diriger des recherches en 1999

à

l'Université de la Méditerranée
Faculté des Sciences de Luminy

Modélisation moléculaire par homologie des protéines : ses applications en Biologie et en Bioinformatique

par

Jean-Luc Pellequer

Soutenance : 22 Janvier 1999
Composition du Jury :
Prof. Michel Fougereau (Président)
Dr. Laurent Chiche (Rapporteur)
Prof. Gilbert Deléage (Rapporteur)
Dr. Jean-Pierre Samama (Rapporteur)
Dr. Jean-Michel Claverie
Prof. Pierre Parot

REMERCIEMENTS

Je voudrais exprimer ma gratitude envers le Jury pour son indulgence concernant ce mémoire du fait des neuf heures de décalage horaire qui séparent Marseille de l'ordinateur où ce mémoire fut écrit.

Je voudrais exprimer ma reconnaissance au Professeur Fougereau pour avoir accepter de présider ce Jury. Je remercie sincèrement les rapporteurs : Dr. Chiche, Prof. Deléage et Dr. Samama pour leurs patiences et leurs commentaires. Enfin, je remercie Dr. Claverie pour sa participation en étayant la compétence du Jury dans le domaine de la bioinformatique et le Prof. Parot pour son soutien logistique à cette entreprise.

Je ne voudrais pas oublier de remercier mon collègue, le Dr. Georges Mer pour son aide précieuse lors de la composition de ce mémoire, en langue étrangère et au pays des claviers sans accents.

Finalement, une pensée chaleureuse à mon épouse Wendy pour sa patience durant cette période où le jour et la nuit ne furent qu'un.

LISTE DES FIGURES

AVANT-PROPOS

Mes activités de recherche se focalisent sur l'étude structurale de la fonction des protéines tant au point de vue expérimental que théorique et plus particulièrement l'aspect de la reconnaissance intermoléculaire. Notre système de référence est celui de la reconnaissance antigène-anticorps. Ce système est le prototype parfait, de part sa fonction, mais également grâce à la dynamique de ce processus lié au concept de maturation. Au cours de ma thèse de doctorat, nous avons mesuré les constantes d'affinité, à l'équilibre et en cinétique, entre des anticorps monoclonaux et un antigène multivalent (virus de la mosaïque du tabac, VMT). De manière inattendue, nos résultats ont démontré une coopérativité négative dans la liaison d'anticorps au VMT. A partir de mon premier stage postdoctoral et jusqu'à aujourd'hui, ma recherche s'est orientée vers l'aspect structural de la reconnaissance ; incluant le développement de fonctions énergétiques et de techniques d'analyse des structures tridimensionnelles, la construction de modèles moléculaires et l'assemblage de ligand dans leur recepteur.

Ce mémoire résume mes cinq dernières années d'activités scientifiques dans le domaine de la modélisation moléculaire, bien que la parution de publications ne reflète guère ce laps de temps. La modélisation moléculaire est une discipline récente qui nécessite, à mon sens, une introduction formelle. Dans ce mémoire on définit la modélisation moléculaire comme l'ensemble des techniques qui permettent d'étudier la fonction d'une molécule grâce à la connaissance de sa structure tridimensionnelle. Ces techniques incluent la modélisation par homologie, les méthodes de simulations, les méthodes d'assemblage (docking), les méthodes d'étude du repliement *ab-initio* des protéines. Les approches spectroscopiques comme la résonance magnétique nucléaire (RMN) ou le dichroisme circulaire et la microscopie électronique (ME) sont également inclues. Au sens littéral notre définition intègre également la diffraction des rayons X (RX).

L'objectif de ce mémoire est de décrire en détail une de ces techniques : la modélisation par homologie. Ce mémoire focalise principalement sur les protéines bien que la modélisation moléculaire s'applique aussi bien aux acides nucléiques, aux sucres ou aux lipides. En dépit de la jeunesse de cette discipline, il est pratiquement impossible de la couvrir en son intégralité. Malgré une recherche bibliographique approfondie, elle est certainement incomplète due à l'interdisciplinarité de cette technique qui couvre tous les champs de recherches, de la théorie fondamentale à la médecine. Ce mémoire espère présenter l'impact de la modélisation par homologie dans la biologie moderne de la manière la plus juste.

Ce mémoire comprend quatre parties. Après une introduction générale, on présentera en détail les diverses approches employées pour la modélisation par homologie. Les contributions personnelles à cette discipline seront enfin exposées. La perspective d'évolution de la modélisation moléculaire sera finalement brièvement discutée.

INTRODUCTION

L'utilisation d'informations tridimensionnelles en biologie est en plein essor. Que ce soit en biologie moléculaire, cellulaire ou en immunologie, tous bénéficient de l'apport d'informations structurales. La modélisation moléculaire créée un lien entre le monde expérimental et le monde structural en produisant des modèles tridimensionnels, en les analysant et en les exploitant dans le cadre de leurs fonctions biologiques (revues : Siezen et al., 1991 ; Bajorath et al., 1993 ; Johnson et al., 1994 ; Eisenhaber et al., 1995 ; Rost & Sander, 1996). Les structures tridimensionnelles permettent de développer de manière rationnelle les expériences nécessaires à l'étude efficace d'une protéine. Comme toutes approches scientifiques, la modélisation moléculaire est un processus cyclique qui propose des hypothèses (sur une base structurale dans notre cas), lesquelles sont testées expérimentalement par le biologiste, et les résultats nourrissent la modélisation en raffinant ces techniques. Nous allons présenté la technique de modélisation par homologie, une technique clée au sein de la modélisation moléculaire.

1. Pourquoi a-t'on besoin de modéliser par homologie ?

La modélisation par homologie trouve sa raison d'être d'une part due aux limites rencontrées avec les techniques expérimentales (RX, RMN, ME) et d'autre part due à l'incapacité actuelle à prédire le repliement d'une protéine *ab-initio* (uniquement par sa séquence).

La technique de diffraction des rayons X, principale productrice de structures tridimensionnelles, est limitée par la nécessité de cristalliser les molécules. Certaines molécules sont instables en solution (protéines transmembranaires), d'autres sont instables en absence de leurs partenaires biologiques. Enfin, la compacité de l'empilement cristallin produit de fortes contraintes sur la conformation de certaines régions localisées à la surface des protéines.
La technique RMN est bien adaptée à l'analyse dynamique d'une protéine en solution mais son utilisation reste encore fortement limitée par la taille des molécules. Cependant des techniques prometteuses pour déplacer ces limites émergent à l'heure actuelle (comme la technique d'alignement permettant la mesure des couplages dipolaires en présence de bicelles ou de particules de virus en bâtonnet- VMT). Finalement, bien que la résolution de la microscopie électronique soit impréssionante, environ 7Å, elle est principalement réservée aux grosses particules comme les virus ou le ribosome qui sont des échantillons moins fragiles que les protéines.

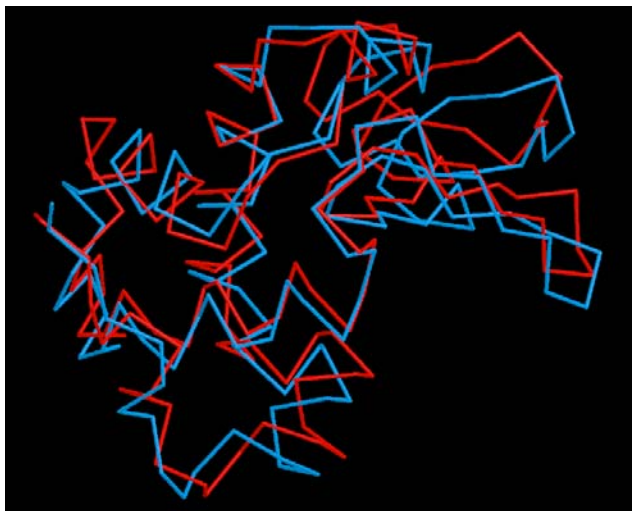
Malgré des efforts considérables, les méthodes de prédiction de repliement peptidique *ab-initio* n'en sont qu'à leur enfance. En l'absence du code de repliement, s'il existe, la

modélisation *ab-initio* requiert deux étapes : la première génère un grand nombre de conformations probables, la seconde discrimine ces conformations grâce à une fonction énergétique. Il est clair qu'il est impossible de générer de manière systématique tous les repliements possibles d'une protéine. En effet, si on considère une des plus petites molécules actives, l'hormone GRH (10 résidus), et que l'on accorde trois conformations pour la chaîne principale ainsi que trois conformations pour chaque chaîne latérale, une recherche exhaustive de la conformation de cette hormone requiert 9^{10} structures à analyser (31.10^9). L'échantillonnage semble être le goulet d'étranglement des techniques *ab-initio*. Plusieurs fonctions énergétiques ont été proposées et bien que d'excellents résultats sont obtenus ces fonctions sont difficilement généralisables à toutes les protéines (Park & Levitt, 1996 ; Park et al., 1997). A l'heure actuelle, les techniques *ab-initio* sont seulement capables de proposer des repliements ayant une déviation d'écart quadratique moyenne (RMSD) de l'ordre de 5 à 15 Å (Defay & Cohen, 1995 ; Zemla et al., 1997).

2. Les fondements de base de la modélisation par homologie.

Le dogme de la modélisation par homologie est basé sur l'observation que les séquences des protéines évoluent plus rapidement que leurs structures et réciproquement que la structure tridimensionnelle d'une protéine est plus conservée que sa séquence.

Il est très probable que les travaux de Perutz ont été décisifs en montrant que la famille des globines avait un repliement tridimensionnel semblable bien que la séquence de la myoglobine diffère singulièrement de celle de l'hémoglobine (Perutz et al., 1965). Bien plus important encore fut la remarque que le repliement des protéines a des caractéristiques communes comme l'exclusion presque totale de résidus polaires à l'intérieur de la chaîne (comme prédit par Kauzmann avant même la connaissance d'une structure protéique). A la vue de cette similarité du coeur hydrophobe des globines, Perutz suggéra qu'il devrait être possible d'identifier des protéines ayant un repliement similaire à l'aide d'un "pattern" de séquences hydrophobes (Perutz et al., 1965). Cette idée de "pattern" s'est transformée en structure primaire et permit au groupe de Phillips de proposer le premier modèle moléculaire de l'α-lactalbumin grâce à son homologie de séquence avec le lysozyme (figure 1, Browne et al., 1969).



Superposition entre l'α-lactalbumin (bleu) et le lysozyme (rouge) illustrant l'homologie structurale en ces deux protéines. Cette figure illustre également la notion de repliement similaire. Lorsque deux protéines ont des repliements similaires, cela ne signifie pas que les deux molécules soient strictement superposables. C'est l'une des grandes limites actuelles de la modélisation moléculaire, i.e. la modélisation ne génère que des repliements connus. La superposition est produite en utilisant les résidus (5-12, 21-36, 78-98, 101-105) pour l'α-lactalbumin et les résidus (5-12, 23-38, 81-101, 111-115) pour le lysozyme. Figure construite par Molscript et Raster3D.

Figure 1: Superposition entre l'α-lactalbumin et le lysozyme.

La notion de famille homogène des protéines s'est propagée chez les protéases à sérines (>45% de résidus identiques) et permit à Hartley de construire deux modèles moléculaires, la trypsine et l'elastase, à partir de la chymotrypsine (Hartley, 1970), puis chez les cytochromes (Almassy & Dickerson, 1978) et les globines (Lesk & Chothia, 1980). En 1981, Doolittle invoque un contexte évolutif à ces homologies structurales en proposant que la plupart des protéines ont évolué à partir d'un faible nombre de protéines archétypes en se basant sur la notion qu'il est plus simple de dupliquer et de modifier des protéines au niveau génomique qu'il est de les assembler *in-vivo* par une combinaison appropriée d'acides aminés (Doolittle, 1981). Si Darwin avait su ?

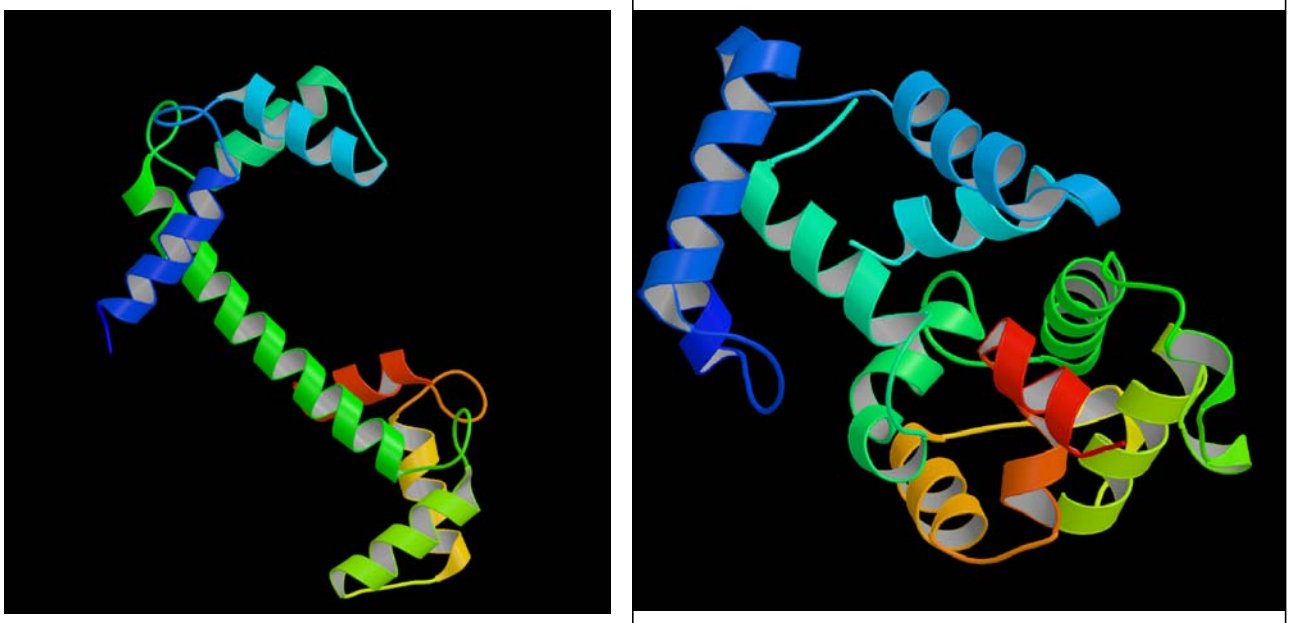
Après ces considérations qualitatives, la modélisation par homologie s'est forgée d'arguments quantitatifs. A partir de l'analyse de huit familles de protéines (Chothia & Lesk, 1986), Chothia a montré qu'une relation mathématique relie la divergence de séquence à la divergence structurale : RMS = 0.40 $e^{1.87H}$ où H est la fraction de résidus mutés (e.g., deux séquences ayant 50% de résidus identiques, H=0.5, auraient une déviation de leur conformation d'environ 1Å). Il définit le coeur des protéines comme étant composé de tous les résidus ayant un RMS < 3Å, et montre que pour des protéines qui ont 50% de résidus identiques, un tel coeur est composé d'environ 90% des résidus, alors que pour des protéines ayant seulement 20% de résidus identiques, le coeur peut inclure entre 42 à 98% des résidus. Il est donc clair que lorsque l'homologie de séquence diminue, l'homologie structurale ne décroît pas; par contre la probabilité de divergence augmente (Chothia & Lesk, 1986). Les premiers modèles moléculaires ont permis d'observer que les variations structurales entre protéines homologues se situent dans les boucles en

surface (Hartley, 1970 ; McLachlan & Shotton, 1971). Une analyse statistique de ces variations structurales (insertions/suppressions ou "insups") indiquent qu'elles sont généralement courtes (de 1 à 5 résidus en moyenne, 99% ayant une taille < 10 residus) et sont localisées dans les tournants ou les boucles (Pascarella & Argos, 1992). La présence d'"insups" au sein des éléments de structures secondaires est un événement rare. D'autres analyses statistiques confirment cette corrélation linéaire entre l'homologie de séquence et l'homologie de structure (protéines ayant <50% de résidus identiques, Flores et al., 1993 ; Hilbert et al., 1993). Un point important concerne la conservation de la conformation des chaînes latérales entre protéines homologues. Il a été montré que l'angle dièdre $\chi 1$ des chaînes latérales enfouies (<15% de surface accessible au solvant- ASA), est conservé à 95% chez des résidus identiques quelque soit le degré d'homologie des protéines homologues, et environ à 60% pour les chaînes latérales ayant une surface accessible au solvant supérieure à 15% (Flores et al., 1993).
En résumé, il a été établi que les protéines ayant une forte homologie de séquence avaient des structures tridimensionnelles identiques et que l'homologie structurale est plus conservée dans le coeur des protéines qu'en surface.

3. Relations séquence-structure et structure-fonction.

Comme nous venons de le voir, le dogme de la modélisation par homologie est que si deux séquences sont similaires alors leurs structures sont proches. Qu'en est-il réellement ? Une forte identité de séquence implique t-elle une identité structurale ?

Doolittle a montré qu'une identité de séquence supérieure à 20% révèle presque toujours une relation structurale (Doolittle, 1981). Une exception connue concerne le repliement tridimensionnel des protéines liant des atomes de calcium (figure 2): calmodulin, sarcoplasmic CBP et recovérine, qui varie de l'ordre de 10Å alors que leurs séquences sont identiques à 30% (Pawlowski et al., 1996). Cependant, il est clair que la présence ou l'absence d'atomes de calcium est un facteur important dans la conformation de ces protéines.



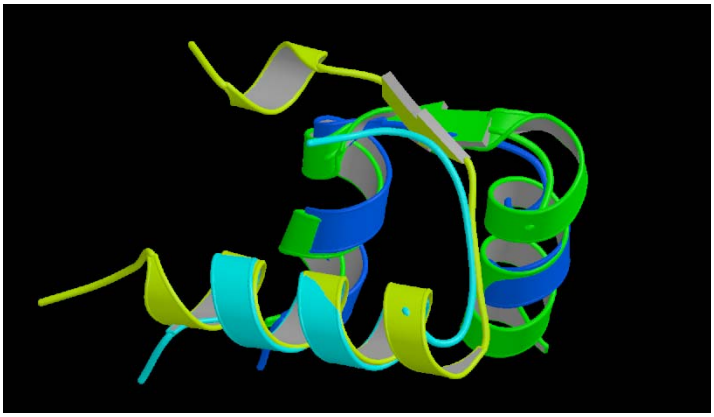
A gauche la calmoduline et à droite la recovérine. Pour mieux comparer leurs structures, elles sont colorées par un gradient de bleu foncé vers rouge (N -> C terminal). La principale différence dans le repliement de la recoverine concerne l'hélice verte centrale qui est coupée en deux comparée à la calmoduline. Figure construite par Molscript et Raster3D.

Figure 2: Exemple de repliements protéiques différents avec des séquences similaires

Une autre exception concerne les petits fragments peptidiques. Il a été montré que des fragments ayant des séquences identiques dans différentes protéines avaient des structures très différentes (Wilson et al., 1985). Ceci fut démontré pour des peptides de cinq (Kabsch & Sander, 1984), six (Cohen et al., 1993) et huit (Sudarsanam, 1998) résidus. Une expérience a même montré qu'un peptide de 11 résidus pouvait adopter une conformation soit en hélice soit en feuillet en fonction de la position à laquelle ce peptide est introduit dans la protéine G (Minor & Kim, 1996). En dehors des considérations importantes pour la compréhension du repliement des protéines, ces résultats indiquent que l'identité de séquence dépend de la taille des fragments et que la modélisation par recherche de petits fragments identiques en séquence peut s'avérer infructueuse. Une formule mathématique décrit l'identité requise en fonction de la taille (L) pour être significative statistiquement : $m = 31L^{-0.124} \pm 18.2\ L^{-0.305}$ où m est l'identité de séquence (Abagyan & Batalov, 1997).

Faut-il avoir une forte homologie de séquence pour être homologue structuralement?
On connaît un grand nombre de protéines qui n'ont aucune homologie de séquence ap-

parente mais qui possèdent un repliement similaire (plasma albumin et α-fetoprotein, β-thromboglobulin et platelet factor, insulin et relaxin figure 3, parvalbumin et troponin C, plastocyanin et azurin, proinsulin et nerve growth factor, serine protease et haptoglobulin, ovalbumin et anti-thrombin III, actin et heat shock cognate protein).



Superposition entre l'insuline en bleu (la chaîne A en foncé, la chaîne B en clair où l'extrémité C terminale est omise pour plus de clarté) et la relaxine en vert. La superposition est produite à partir des résidus A3-A8, A13-A17, B10-B19. Figure construite par Molscript et Raster3D.

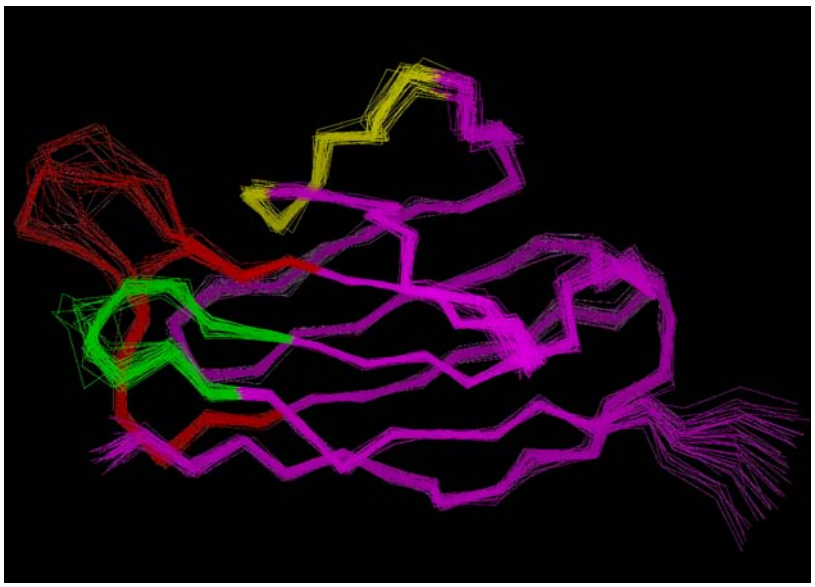
Figure 3: Exemple de similarité structurale en absence de similarité de séquences.

Si des protéines éloignées en séquence ont le même repliement, combien y a t-il de repliement en tout ?
Il est estimé qu'il y a entre 1000 et 7000 éxons différents, avec une taille moyenne de 40 à 50 résidus (Dorit et al., 1990). Cependant la relation entre éxon et repliement n'est pas démontrée. Chothia estima qu'il y avait environ 1000 familles de protéines (Chothia, 1992), sous-entendu 1000 repliements différents de protéines. Le nombre de repliement maximum fut estimé à 8000 (Orengo et al., 1994), mais ce chiffre est largement constesté par une récente analyse qui prédit environ 800 repliements différents (Zhang & DeLisi, 1998). Ces estimations sont basées principalement sur le fait qu'environ 35% des nouvelles molécules cristallisées ont des repliements déjà connus (Orengo et al., 1994 ; Holm & Sander, 1997). La difficulté d'une estimation précise est liée aux grandes protéines (>600 résidus) qui souvent sont composées de plusieurs domaines dont les repliements ont déjà été observés.

Une homologie structurale implique t-elle une homologie fonctionnelle ?
Cette question est centrale aux préoccupations des bioinformaticiens qui analysent les séquences générées par le séquençage des génomes. Bien que dans la majorité des cas, une homologie structurale implique une homologie fonctionnelle des exceptions existent. Par exemple, l' α-lactalbumin à une structure identique à celle du lysozyme, mais en a perdue son activité catalytique ; ou encore les protéines ayant le repliement de tonneaux α/β possèdent des fonctions très variées (isomérase, kinase, aldolase, oxidase, deshy-

drogénase...).

Quelle précision peut-on espérer atteindre avec un modèle moléculaire ?
Il est clair que plus les séquences de protéines sont divergentes, plus leurs structures le seront (Chelvanayagam et al., 1994 ; Russell & Barton, 1994 ; Chung & Subbiah, 1996). Plusieurs formules mathématiques ont été proposées pour relier le degré d'homologie de séquence à la déviation des écarts quadratiques moyens (RMSD) des repliements (e.g., $\Delta = 0.40\ e^{1.87H}$, $S=12.3 + 98.4\ e^{-0.95\Delta}$) où S et H sont l'identité de séquence et Δ le RMSD. La première formule montre qu'une identité de 50% (H=0.5) implique un RMSD d'environ 1Å (Chothia & Lesk, 1986) alors que la seconde montre qu'une déviation de 1Å correspond à une identité de 50% alors qu'une déviation de 3Å correspond à une identité de 18% (Chelvanayagam et al., 1994). Un exemple est donné dans la figure 4.



Superposition d'environ 50 structures de chaînes légères d'immunoglobulines. L'homologie de séquence est environ 80% alors que l'homologie structurale est d'environ 1Å pour le coeur. Figure construite par InsightII.

Figure 4: Variations structurales au sein d'une famille de protéines conservées.

Une autre partie de la reponse à cette question dépend de la résolution des structures cristallographiques, en terme expérimental (facteur d'accord R), et en terme fonctionnel. Autrement dit, existe t-il une structure unique pour chaque molécule ? La réponse est négative. Une expérience intéressante a montré qu'un réaffinement de quatre structures cristallographiques de la même molécule par un jeu unique de données résultent en une divergence de l'ordre de 0.84Å pour tous les atomes (Ohlendorf, 1994), bien plus que prévu par le critère classique de Luzzati (0.2-0.4Å). On peut donc se poser la ques-

tion : comment représenter la structure d'une protéine ?
De plus, les structures expérimentales ne sont pas sans incertitudes. Par exemple, une structure cristallographique ayant un facteur d'accord de 20% est considérée comme acceptable. Ce facteur indique qu'environ 20% des données expérimentales ne peuvent être reproduites par le modèle tridimensionnel proposé. De plus, il existe bien souvent des régions sans densité provoquant des coupures dans la chaîne peptidique. Tous ces facteurs ont un impact certain sur la qualité d'un modèle moléculaire.
En résumé, les protéines sont des molécules flexibles surtout dans les régions exposées au solvant mais aussi dans les régions internes comme le montre les expériences de RMN. Une forte homologie de séquence indique une probable similarité structurale et fonctionnelle.

Quelles sont les limites de la modélisation ?
Il est trivial que la limite principale de la modélisation par homologie est le manque de structures tridimensionnelles homologues. Sans structures parentes expérimentales (RX ou RMN), la modélisation n'a pas de raison d'être. De plus certaines molécules sont peu représentées dans les bases de données. Ceci est sans doute lié au biais naturel envers la résolution de structures tridimensionnelles dû : à leurs importances biologiques, à la faciliter d'extraction, de purification et de production, à leurs stabilités en solution et à leurs tailles (Bajorath et al., 1993 ; Johnson et al., 1994).
Une autre limite à l'heure actuelle est la difficulté, voir l'impossibilité, de prédire les changements conformationnels qui se produisent durant une activité biologique. En effet, il est très fréquent d'observer des changements conformationnels de type "charnières" (hinge) où l'orientation de deux domaines structuraux (ou deux éléments de structures secondaires) d'une protéine peut varier l'un part rapport à l'autre. En particulier, ces changements conformationels sont fréquents dans les protéines liant des ligands, notamment les immunoglobulines. Toutefois, cette limite dépend plus de notre manque de connaissance dans l'analyse des structures tridimentionnelles que dans les méthodes de modélisations.

4. Stabilité et repliement des protéines.

La compréhension des forces responsables de la stabilité des protéines est importante dans le domaine de la modélisation moléculaire. Ces forces permettent d'estimer la validité d'un modèle moléculaire, de comprendre l'insolubilité des protéines en milieu aqueux, de prédire la stabilité d'une mutation ou d'estimer l'énergie d'interaction entre deux molécules.

Les protéines sont toutes construites sur le même moule qui se caractérise par un enfouissement partiel de sa surface accessible (proportionnelle à sa masse moléculaire

$A=0.859M + 0.774M^{2.10\text{^-}5}$) en incluant une proportion constante de groupes polaires, ainsi que par une grande compacité du coeur enfoui (Chothia, 1975). La stabilité des protéines est marginale vis-à-vis de l'état dénaturé avec un excédant de l'ordre de -5 à -10 kcal/mol (Pace et al., 1996). Le repliement des protéines est gouverné par leur séquence et elles se situent dans leurs minimums énergétiques conformationnels (Anfinsen, 1973).

Il y a principalement cinq forces responsables de la stabilité des protéines (Kauzmann, 1959) : l'effet hydrophobe, la liaison hydrogène, le pont salin, les forces de dispersions (London) et les ponts disulfures. On va décrire en détail l'effet hydrophobe en raison de son importance alors que les quatre autres forces seront brièvement décrites.
On peut expliquer l'effet hydrophobe par la tendance qu'on les groupements non polaires à adhérer les uns aux autres dans un environnement aqueux. En d'autres termes, l'énergie de transfert d'un groupe non polaire dans l'eau est > 0 (□θ pour des groupes aromatiques) avec une composante enthalpique presque nulle et une forte composante entropique ($\Delta S<0$). Comme le montre Kauzmann, les forces qui contrôle le repliement sont de nature entropique et concernent principalement la réorganisation du solvant autour d'une protéine (il s'agit de la théorie des "icebergs" où lorsqu'une molécule non polaire est présente dans un milieu aqueux, les molécules d'eau à l'interface doivent s'arranger dans une forme quasi-cristalline provoquant une diminution de l'orientation aléatoire des molécules d'eau et un accroissement de la force des liaisons hydrogène comparée à celle de l'eau ordinaire à la même température). Lors de la dénaturation d'une protéine, la formation "d'icebergs" augmente parce que les "liaisons hydrophobes" sont cassées et exposées au solvant (Kauzmann, 1959). Le coût entropique pour geler l'eau en glace est de l'ordre de 5.3 cal/deg.mol alors que la perte d'entropie lorsqu'une molécule non polaire entre dans l'eau est d'environ 20 cal/deg.mol. Par conséquent il est estimé qu'un faible nombre de molécules d'eau participe au processus. En ce qui concerne la stabilité des protéines, l'enfouissement de groupes non polaires dans un coeur de protéines est donc énergétiquement très favorable.
La liaison hydrogène est un sujet très controversé tant au niveau énergétique qu'au niveau de son importance dans le processus de repliement des protéines. Kauzmann indique que les liaisons hydrogène stabilisent les protéines par -0.4 kcal/mol à 25degré alors que Fersht propose une gamme allant de -0.5 à -1.8 kcal/mol (Fersht et al., 1985) et Pace estime une valeur de -1.5 kcal/mol grâce aux résultats de mutagénèses dirigées (Asn->Ala, Ser->Ala, Val->Thr, Tyr->Phe, Pace et al., 1996). L'importance des liaisons hydrogène est apparente au travers de ces quelques chiffres : au cours du repliement, une protéine enfouie 81% des chaînes latérales non polaires, 70% des groupes de la chaîne principale, 63% des chaînes latérales polaires et 54% des chaînes latérales chargées (incluant les groupements méthylènes, Lesser & Rose, 1990). Il est donc clair que l'intérieur des protéines n'est pas purement hydrophobe et que la balance électrosta-

tique des forces impliquées est très complexe. Cette polarisabilité de l'intérieur des protéines a permis à plusieurs groupes de postuler une forte constante diélectrique (de 2 à 20) dans le but de prendre en compte la réorganisation des charges.
La stabilité produite par les ponts salins est également en débat. Un point d'accord semble être que l'apport énergétique d'un pont salin est presque nul et que la formation d'un pont salin est sous contrôle entropique (Kauzmann, 1959). En effet, lorsque deux charges sont séparées dans un milieu aqueux, elles sont environnées par des molécules d'eau fortement orientées de par les forces de compression provenant du champ électrique au voisinage des charges. Lorsque ces charges viennent en contact, le champ électrique passe de moins en moins bien à travers les molécules d'eau et le solvant devient moins orienté et moins compressé.
Les forces de dispersions (London) sont des forces toujours attractives qui permettent à des groupes identiques de se réunir. Ces forces sont souvent représentées par une contribution énergétique variant à l'inverse de la puissance 6 de la distance entre deux atomes.
Les ponts disulfures introduisent une pénalité entropique non négligeable puisqu'il est estimé que la rigidification d'une chaîne de longueur N décroît par un factor 3/2 RlnN. Par exemple, un pont disulfure entre 100 résidus réduit l'entropie de +4.1 kcal/mol pour le repliement de la molécule native (Kauzmann, 1959).

Malgré la stabilité marginale des protéines, une grande quantité de résultats expérimentaux montrent que les protéines sont relativement résistantes aux mutations (Bowie et al., 1990) aussi stringentes que l'introduction de résidus chargés dans le coeur hydrophobe de la thioredoxine (Hellinga et al., 1992) ou encore des mutations introduites en plein milieu d'une hélice du lysozyme T4 (Heinz et al., 1993). Cette stabilité, principalement liée à la flexibilité d'adaptation de la structure tridimensionnelle, a toutefois une limite comme le montre les expériences de mutagénèses intensives sur le lysozyme T4 où parmi les 2015 mutations, celles affectant les résidus enfouis, hydrophobes ou chargés, ou les résidus fonctionnels ont un effet délétère (Rennell et al., 1991). De plus, il a été montré que les résidus conservés au travers d'une famille sont beaucoup plus sensibles (100% des 14 résidus conservés du lysozyme T4) que les autres à la mutagénèse (Poteete et al., 1992).

5. Les grandes applications de la modélisation par homologie.

Avant de présenter les techniques de modélisation dans leurs détails, il est instructif de regarder vers le passé et d'établir un bilan des grands succès de la modélisation par homologie. Evidemment, la notion de succès est arbitraire : le succès de la modélisation *per se* (la preuve qu'il est possible de prédire la conformation d'une protéine à partir d'une molécule de structure connue), et le succès pratique (nouvelle découverte, production

d'une drogue efficace, compréhension des mécanismes fonctionnels).

De manière inattendue, la modélisation par homologie de la trypsine a révélé une erreur dans sa séquence pour le résidu 189 (Asn) qui ne permit pas de comprendre, au vue de la structure tridimensionnelle, la spécificité de la trypsine pour les résidus positivement chargés. Ce résidu a été remplacé par un acide aspartique (Hartley, 1970). La modélisation par homologie a permis à McLachlan de montrer, contre toute anticipation, que la conformation de l'α-lytic protease était très voisine de celle de la chymotrypsine ou de l'élastase malgré une faible homologie de séquence (McLachlan & Shotton, 1971). Cette similarité a été démontrée ultérieurement grâce à la structure cristalline de l'α-lytic protease (Delbaere et al., 1979).

Les plus grands succès pratiques de la modélisation incluent trois thèmes majeurs : l'ingéniérie d'inhibiteurs ou d'analogues et l'assemblage de ligands dans leurs récepteurs, l'identification de régions potentielles d'interaction, et l'analyse de mutations ponctuelles.
Le premier thème a évidemment plus d'impact que les autres. Une liste d'assemblage de ligands dans leur recepteur est proposée dans l'annexe II. En ce qui concerne le développement d'inhibiteurs, des succès marquant ont été obtenus pour la dihydrofolate reductase (Hansch et al., 1982 ; Kuyper et al., 1985), prealbumine (Blaney et al., 1982), anhydrase carbonique (Baldwin et al., 1989), thymidilate synthase (Appelt et al., 1991 ; Varney et al., 1992 ; Shoichet et al., 1993), purine nucleoside phosphorylase (Ealick et al., 1991 ; Erion et al., 1993 ; Montgomery et al., 1993 ; Secrist III et al., 1993), protéase à serine ou à cystéine (Ring et al., 1993), replication du virus Influenza (von Itzstein et al., 1993), thrombine (Hilpert et al., 1994), la protéase du VIH (Lam et al., 1994), interleukine 2 (Tilley et al., 1997) ou encore le recepteur à tyrosine kinase erbB (Singh et al., 1997).
Le second thème est important car il est source de génération d'un grand nombre d'hypothèses. Par exemple, des sites d'interaction ont été proposés pour : un peptide basique sur la calmoduline (O'Neil & DeGrado, 1985), le centre phostosynthétique II (Svensson et al., 1990), l'ATP dans une kinase (Schoentgen et al., 1992), un épitope linéaire pour la protéine C4b (Fernández et al., 1994), l'héparine dans une protéase (Matsumoto et al., 1995), l'hormone LH/CG à son recepteur (Bhowmick et al., 1996), la prostaglandine H2 à son recepteur (Wang et al., 1996), le site de liaison à l'ADN d'une synthétase (Sticht et al., 1997), Arp et l'héparine sur C4b (Villoutreix & Dahlback, 1998). Pour les modèles d'anticorps il est également possible de prédire les résidus impliqués dans le site de liaison (Ruff-Jamison & Glenney, 1993).
Le troisième thème propose des explications structurales ou énergétiques pour certaines mutations. Par exemple, la définition de la sévérité des mutations dans la cascade de coagulation (Greengard et al., 1994), les mutations non-sens dans les malades atteints de

galactosialidosis (Elsliger & Potier, 1994), les mutations sur le facteur de coagulation humain VIII responsable de l'hémophilie (Pan et al., 1995 ; Pemberton et al., 1997), mutations dans la maladie methylmalonic aciduria (Thomä & Leadlay, 1996), sur le recepteur androgène humain (Lobaccaro et al., 1996) ou sur une phosphatase (Quondam et al., 1997).

La modélisation par homologie a également permis d'étudier : le mécanisme de transfert de signal dans le recepteur oestrogène humain (Lewis et al., 1995), la stabilité intrinsèque d'une protéine à atomes de cuivre (Grossmann et al., 1995), la thermostabilité (Szilágyi & Závodszky, 1995), le fonctionnement du blocage d'un canal ionique avec l'amantadine (Samsom & Kerr, 1993), l'assemblage supra-moléculaire de l'α-tropomyosin (Cregut et al., 1993), la stéréochimie de la sélectivité d'un inhibiteur de la dihydrofolate reductase (Matthews et al., 1985) ou enfin de développer un ligand permettant de lier la cyclophiline à la calcineurine *in-vivo* (Alberg & Schreiber, 1993).

Un domaine de recherche nouveau s'appuie considérablement sur les techniques de la modélisation moléculaire. Il s'agit de l'ingéniérie des protéines par voie de design rationel. L'ingéniérie consiste à développer une nouvelle fonction protéique à partir de l'architecture d'une protéine connue. Par exemple, la création de protéines captrices de métaux (Hellinga & Richards, 1991 ; Gregory et al., 1993 ; Hornischer & Blöcker, 1996), la grèffe d'un fragment peptidique inter-protéine (Hearst & Cohen, 1994), l'introduction de ponts disulfures (Pabo & Suchanek, 1986), l'humanisation des anticorps de souris (Kettleborough et al., 1991 ; Riechmann et al., 1992 ; Presta et al., 1993), l'ingéniérie de l'α-hordothionin pour augmenter son contenu en lysine dans le but d'équilibrer l'apport nutritif des alimentations végétariennes (Rao et al., 1994), le changement de la spécificité d'une protéase à sérine en remplacant une arginine par une glycine (Caputo et al., 1994) ou encore l'amélioration de la stabilité et l'expression d'un fragment variable d'anticorps simple chaîne (Nieba et al., 1997)

Ces applications de la modélisation soulignent la pluridisciplinarité de cette technique. L'objectif de cette liste, certes partielle, est de démontrer que bien que récente, la modélisation par homologie a obtenu des succès probants et que son utilisation va devenir de plus en plus impliquée dans la recherche biologique.

MODELISATION PAR HOMOLOGIE

La modélisation par homologie comprend six étapes incontournables, certaines plus ou moins triviales que d'autres dépendant de la famille de protéine étudiée (Swindells & Thornton, 1991 ; Sali, 1995) :

1) Recherche d'un repliement connu (parent)
2) Alignement des séquences (cible et parent)
3) Construction de la chaîne principale
4) Contruction des chaînes latérales
5) Constructions des insertions/suppressions ("insups")
6) Optimisation et affinement du modèle

1. Recherche d'un repliement connu.

Lorsque l'homologie de séquence est significative, la recherche d'un repliement dans la Protein Data Bank (PDB) est généralement triviale avec le programme BLAST (Altschul et al., 1990) accessible sur le site web du NCBI (www.ncbi.nlm.nih.gov). L'avantage d'utiliser le programme BLAST est qu'il fournit un alignement des séquences avec des espaces (GAP) dans les régions les moins conservées, et plus important, un score statistique de la significativité de la relation entre les deux séquences alignées. Cependant, dans de nombreuses situations, les alignements proposés sont au-delà du seuil statistique de significativité en raison d'une grande divergence de séquence. Comme nous l'avons vu auparavant, la faible homologie de séquence n'empêche pas la similarité structurale.
Dans le but d'établir un lien entre l'homologie de séquence et de structure pour des molécules divergentes, l'utilisation d'une technique de "threading" est envisageable (figure 5). Le concept du threading est de mesurer l'adéquation entre un repliement tridimensionnel et une séquence peptidique. Que se soit par matrices 3D-1D (Bowie et al., 1990 ; Fischer & Eisenberg, 1996) ou par une fonction statistique qui représente la probabilité qu'un résidu soit au voisinage d'un autre (Sippl, 1990 ; Jones et al., 1992), cette technique a permis à plusieurs reprises d'identifier des homologies structurales même lorsque l'identité de séquence est inferieure à 20% (Wodak & Rooman, 1993 ; Kocher et al., 1994).

RANK	Z-SCORE	FOLD	LENGTHALI	%ID	
1	5.82	1gof_1-150	130	20	; Galactose-binding domain-like
2	4.0	1pex	135	27	; COLLAGENASE-3 (MMP-13)
3	3.82	1bak	107	18	;
4	3.32	1pgs_4-140	123	18	; Glycosyl-asparaginase
5	3.04	1cd1a	144	25	; CD1(MOUSE) ANTIGEN PRESENTING MOL.
6	3.03	1cmba	104	18	; Met repressor-like
7	2.80	1pls	99	23	; PH domain-like] 2-3

Exemple de repliements identifiés pour le domaine C2 du facteur de coagulation Va avec un programme de threading (3D-1D). Le test statistique (Z-score) indique que la seule reponse satistiquement significative est la première (1gof_1-150, >4.8) avec seulement 20% de résidus identiques sur une longueur de 130 résidus.

Figure 5: Résultats du "threading" sur le domaine C2 du factor Va.

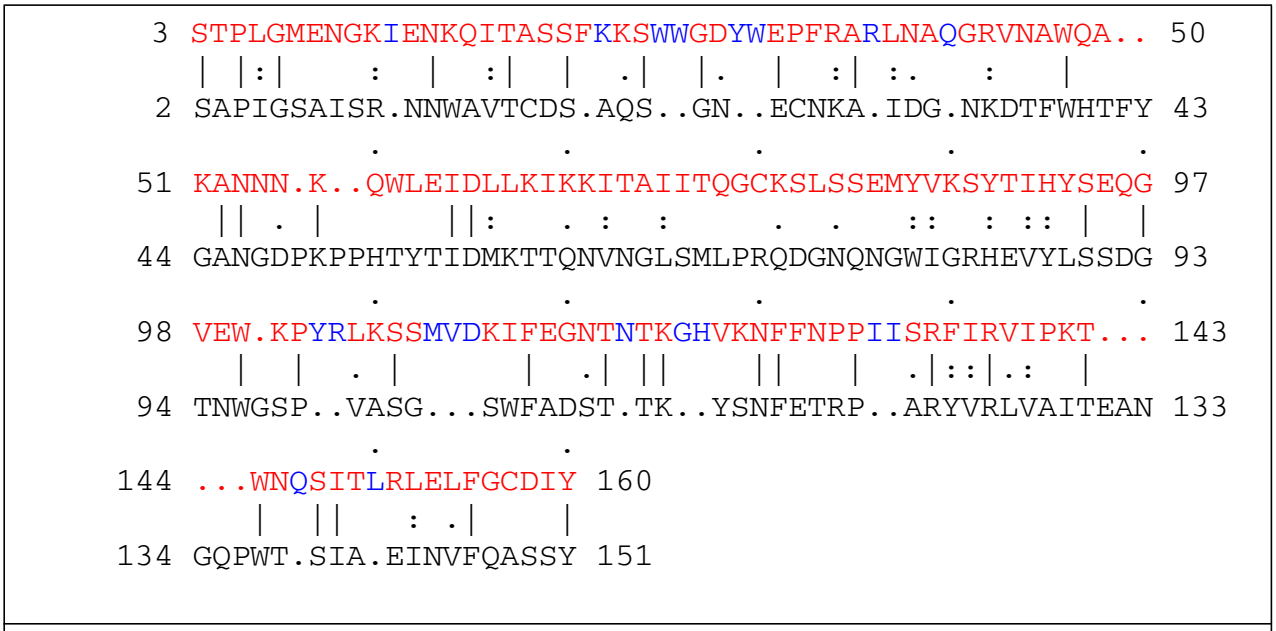
En cas de grande divergence l'utilisation de motif peut s'avérer cruciale. Un motif peut être considérer comme une signature commune entre une famille de protéines (Bairoch, 1991 ; Han & Baker, 1996 ; Hutschinson & Thornton, 1996). Bien que la création de profil soit souvent manuelle, les résultats obtenus peuvent être révélateurs (Pellequer et al., 1998).

2. Alignement de séquences.

Conceptuellement trivial, l'alignement de la séquence cible avec la séquence parente est l'opération la plus difficile et la plus délicate du processus de modélisation par homologie. En effet, une erreur dans l'alignement peut provoquer des déplacements d'éléments de structures secondaires pour des raisons d'incompatibilité de chaînes latérales ou encore insérer une boucle au mauvais endroit dans la chaîne principale. Dans la plupart des cas, la reconnaissance de structure parente adéquate (étape no1) donnera naissance à un alignement provisoire.

Lorsque l'on utilise un programme d'alignement de séquence basé sur la technique récursive de Needleman (Needleman & Wunsch, 1970), le choix de la matrice de similarité est crucial. Il existe plusieurs types de matrices depuis leurs apparitions (Fitch, 1966 ; McLachlan, 1971). La plus célèbre est certainement la matrice PAM (Point Accepted Mutation) basée sur l'observation de mutations de résidus au sein d'une même famille de protéines (Dayhoff, 1969 ; Gonnet et al., 1992). Certaines incluent des données structurales comme la probabilité de trouver un résidu dans une conformation particulière dans la carte de Ramachandran (Risler et al., 1988 ; Niefind & Schomburg, 1991) alors que BLOSUM est contruite à partir de block de séquences (Henikoff & Henikoff, 1992).

Il a été montré dans plusieurs travaux que les matrices BLOSUM et Gonnet donnent les meilleurs résultats (Johnson & Overington, 1993 ; Abagyan & Batalov, 1997). Certaines méthodes proposent également des variantes comme l'incorporation du volume des acides aminés (Saqi et al., 1992) ou l'introduction d'un poids supplémentaires qui autorise des séquences à mal s'aligner (Alexandrov & Leuthy, 1998).
Quoiqu'il en soit, aucune méthode à l'heure actuelle ne propose un alignement idéal entre deux séquences (figure 6).



Alignement de séquences produit grâce à la matrice BLOSUM62 avec des pénalités de "gaps" égales à 4 (création) et 2 (extension) pour le domaine C2 du facteur de coagulation Va (rouge) et le domaine de liaison de la galactose oxidase (noir). Ces faibles poids sont nécessaires pour obtenir un alignement sur toute la longueur des protéines. Il est clair qu'avec 12 insertions (en bleu) et 18 "gaps", cet alignement est inutilisable pour la modélisation par homologie.

Figure 6: Alignement automatique de séquences avec le programme BESTFIT de GCG.

Pourquoi ne pas utiliser le "threading" pour aligner les séquences ?
Deux raisons s'y opposent : la première est qu'il n'y a pas de guarantie pour qu'une reconnaissance de repliement identifié soit réelle, même avec un score significatif (Flöckner et al., 1995), la deuxième est que l'alignement de séquence est loin d'être optimal (figure 7), en particulier au niveau des insertions/suppressions (Hubbard & Park, 1995 ; Jones et al., 1995 ; Di Francesco et al., 1997 ; Rice & Eisenberg, 1997) bien que des progrès récents aient été observés (Flöckner et al., 1997).

Alignement obtenu entre le domaine C2 (rouge) du FVa et le domaine de liaison de la galactose oxidase (noir). Les éléments de structures secondaires sont représentés au-dessus et au dessous de leur séquence respective (pour le domaine C2, se sont les éléments de structures secondaires prédit par le programme PHD). Même si l'alignement indique une identité de séquence de 18%, plusieurs gaps sont localisés dans les éléments de structures secondaires.

Figure 7: Alignement de séquences produit par un programme de "threading".

Comment intégrer les données structurales connues de la molécule parente dans l'alignement ?

Un pas important a été franchis grâce à la définition des régions structuralement conservées (SCR) et des régions variables (VR) par Greer (Greer, 1981a ; Greer, 1990 ; Greer, 1991). Ces définitions sont basées sur l'homologie structurale présente au sein d'une famille de molécules. Lorsque les structures tridimensionnelles de plusieurs molécules de la même famille sont connues, il est possible de les superposer les unes sur les autres et de définir par conséquent les régions conservées et les régions variables grâce à un critère de type RMSD. Le premier programme capable de déterminer les SCRs automatiquement fut COMPOSER qui employa un critère RMSD de 3Å sur les carbones α (Cα, Blundell et al., 1987 ; Sutcliffe et al., 1987 ; Blundell et al., 1988 ; Srinivasan & Blundell, 1993).

Il est également possible de déterminer les SCRs en ce basant sur une définition précise du coeur des protéines. Par exemple, on peut considérer les résidus ayant un pourcentage de surface accessible au solvant inférieure à 7% (Hubbard & Blundell, 1987), ou utiliser un seuil RMSD entre des matrices de distances inter-carbones α (Lee, 1992). La supposition sous-entendue est que les SCRs correspondent aux régions conservées au niveau de la séquence. Une fois les SCRs assignées il est possible d'optimiser l'alignement de la séquence cible en maximisant le recouvrement des séquences dans les régions SCRs. Conceptuellement, la technique des SCRs à deux effets : elle restreint l'introduction d'insertions/suppressions dans le coeur d'une protéine et permet de localiser les régions les plus variables dans les boucles comme site priviligié des "insups".

Malgré toutes ces précautions, des erreurs d'alignement persistent. Un expert averti peut, par visualisation graphique et beaucoup d'intuition, modifier l'alignement pour provoquer le moins de changement possible dans la chaîne principale lors de la modélisation d'"insups", mais beaucoup de travail reste à faire.

3. Construction de la chaîne principale.

Les premiers modèles moléculaires furent construits manuellement par des jeux de constructions (Browne et al., 1969 ; Hartley, 1970 ; McLachlan & Shotton, 1971 ; Kretsinger & Barry, 1975 ; Padlan et al., 1976 ; Bedarkar et al., 1977 ; Blundell et al., 1978 ; Isaacs et al., 1978). La venue d'outils de superpositions (McLachlan, 1972), de visualisation graphique en mode CPK (Corey-Pauling-Koltung, Feldmann et al., 1978) et de modélisation interactive (Jones, 1978 ; Langridge et al., 1981) permit d'accélerer la construction et d'automatiser certaines étapes comme la mesure manuelle d'angles dièdres.
On compte trois méthodologies permettant de construire la chaîne principale. La première, n'étant pas exactement une méthode de modélisation par homologie consiste à construire la chaîne principale à partir de la connaissance de la position des Cα. La seconde prend avantage de la définition des SCRs. La troisième consiste en une reconstruction par détermination de contraintes.

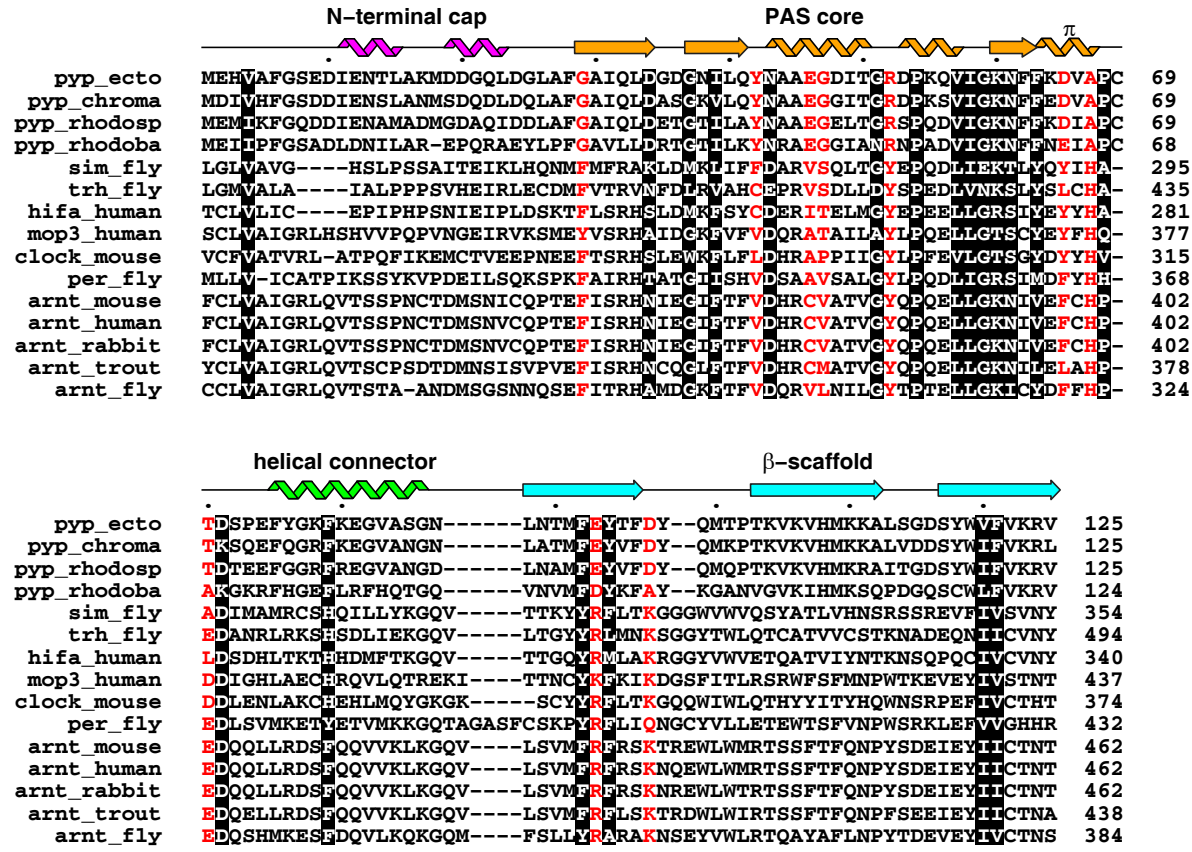
a. A partir de données structurales incomplètes (Cα).

Les constructions de chaînes principales à partir des carbones α ont été très inspiratrices pour la modélisation des boucles. Une vertu de cette technique est de pouvoir tester plusieurs protocoles d'affinement dans un contexte de modélisation.
Un premier concept, très utilisé en cristallographie, consiste à cribler une banque de données de fragments structuraux grâce à des critères de distance entre les Cα. Cette approche fut incorporée dans le programme FRODO dans le but d'accélerer la construction d'un modèle dans la carte de densité électronique (Jones & Thirup, 1986). Une variante consiste à prédéfinir des fragments peptidiques structuraux, d'une taille de six résidus par une méthode de "clustering", et de construire la structure désirée par recouvrement de fragments (Unger et al., 1989).
Le seconde concept consiste à cribler la base de données de structures par une simple superposition des Cα (Claessens et al., 1989). Une variante introduit des critères d'homologie de séquences et énergétiques sur les fragments ainsi obtenus (Levitt, 1992).
Le troisième concept utilise une approche *ab-initio*. La construction progressive où chacun des résidus sont construits un à un, suivit d'une minimisation restreinte, a été proposé par Correa (1990). Lorsque la chaîne principale est complétée, une optimisation par dynamique moléculaire (MD) à haute température est réalisée (1000K). Ensuite les chaînes latérales sont construites par niveaux (Cγ, Cδ, Cε...) entrecoupés de MD à 800K

avec la chaîne principale immobile. Une variante fut proposée, où les atomes Cβ sont construits en premier grâce à une méthode géometrique, suivit par la construction de la chaîne principale et terminée par une minimisation en utilisant des contraintes de distances (Rey & Skolnick, 1992). D'autres méthodes ont été proposées comme la technique de construction grossière (CSB, van Gelder et al., 1994) ou par simulation MC basée sur une grille de probabilité d'angles dièdres (Mathiowetz & Goddard, 1995).

b. Construction employant les SCRs ou la structure secondaire.

La définition des SCRs, outre leur intérêt au niveau de l'alignement de séquence, permet



L'alignement multiple entre les protéines ayant un domaine PAS n'a pu être établi une fois que le programme de prédiction de structures secondaires (PHD) indiqua que le côté Cterminal du domaine PAS inclue trois brins de feuillets comme dans la structure tridimensionnelle de PYP (Pellequer et al., 1998a).

Figure 8: Alignement final entre PYP et des protéines ayant un domaine PAS.

de construire la chaîne principale par simple transposition des coordonnées des SCRs de la structure parente. Lorsque plusieurs structures parentes sont disponibles, deux vari-

antes existent : la première consiste à utiliser les SCRs de la structure ayant la plus grande homologie de séquence, la seconde consiste à mélanger les SCRs provenant de plusieurs structures parentes en fonction de leurs homologies locales de séquences avec la molécule cible. L'annexe II indique la méthode employée par l'acronyme SCR pour la première variante et l'acronyme SCRmix pour la seconde.
Dans le cas où une seule structure parente est connue (souvent la majorité des cas), il est d'usage de considérer que les SCRs correpondent aux éléments de structures secondaires. L'utilisation d'information complémentaire est utile, en particulier la prédiction de structure secondaire lorsque les homologies de séquences sont faibles (figure 8).
Une méthode alternative a été proposée qui, au lieu d'utiliser les SCRs les plus homologues, utilise les SCRs correspondant aux séquences d'éxons (Kajihara et al., 1993). Cependant, le succès de cette méthode est mitigé.

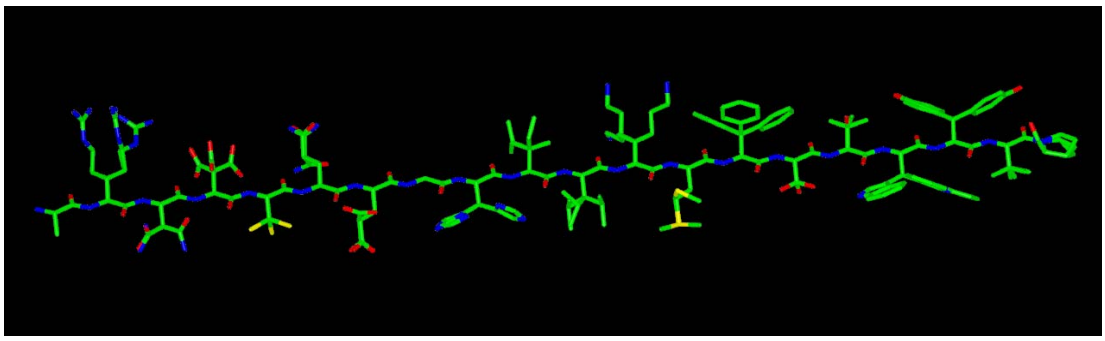
c. Reconstruction par détermination de contraintes.

Le pionnier de cette méthodologie est Havel qui s'intéressa à la reconstruction de protéines grâce à des constraintes de distances obtenues expérimentallement par la technique RMN (Havel et al., 1979). Le premier programme (DISGEO) a permis de construire la première structure RMN de protéine (Havel & W¸thrich, 1985). En 1991, Havel proposa d'utiliser les contraintes de distance pour modéliser de nouvelles protéines à partir d'une ou plusieurs structures parentes. Après avoir déterminé une liste de contraintes entre la distance des atomes, un seuil de confiance proportionnel à l'homologie de séquence entre la proteine cible et la protéine parente est établie sur ces distantes (Havel & Snow, 1991 ; Srinivasan et al., 1993). La grande force de cette technique est de pouvoir moduler le niveau de confiance des distances à volonté (Havel, 1993). Dans cette technique il est souvent nécessaire d'ajouter des contraintes d'angles dièdres ou de liaisons hydrogène et d'optimiser les structures par une technique de recuit simulé (Brocklehurst & Perham, 1993 ; Havel, 1993). L'optimisation de manière progressive des contraintes de structures secondaires suivi de l'optimisation de toutes les contraintes a été également introduit (Fujiyoshi-Yoneda et al., 1991).
Au lieu d'utiliser des contraintes de distances, Sali proposa des fonctions de densité probable (PDF) déduites à partir de contraintes de distances (inter-Cα, N-O, chaîne principale-chaîne latérale). Le programme MODELLER définit automatiquement les PDFs et minimise ces fonctions par la méthode des gradients conjugués (Sali & Blundell, 1993). Une variante dans la détermination des contraintes de distances fut d'autoriser les atomes de la chaîne principale du coeur de dévier de leur position originale dans le but de permettre une meilleure relaxation des boucles lors de la modélisation de protéines divergentes (FOLDER, Sudarsanam et al., 1994). Le programme DRAGON, en plus des contraintes de distances classiques, introduit des contraintes sur les éléments

de structures secondaires améliorant sensiblement la qualité du modèle (Aszódi & Taylor, 1996).

4. Construction des chaînes latérales.

L'analyse de la conformation des chaînes latérales attira rapidement l'attention des chercheurs. Dès 1970, Chandrasekaran montra que les chaînes latérales occupent des positions privilégiées compatibles avec les conformations éclipsées et trans des composés chimiques (Chandrasekaran & Ramachandran, 1970). L'analyse des chaînes latérales dans les protéines révèla que leur conformation est représentative de celles trouvées dans les acides aminés isolés (Gelin & Karplus, 1975) et que la position gauche plus (-60) est la plus favorable (Janin et al., 1978). La conformation des chaînes latérales est principalement gouvernée par des considérations stériques (Bhat et al., 1979) et une certaine dépendance vis-à-vis de la conformation de la chaîne principale est perceptible (Janin et al., 1978).



Représentation du dictionnaire de rotamères présent dans le programme XFIT. Les rotamères proviennent de l'étude de Tufféry et al., 1991. Seuls les trois rotamères les plus favorables sont représentés pour plus de clarté. A gauche se trouve l'arginine, la chaîne se poursuit vers la droite dans l'ordre décroissant de l'alphabet des acides aminés à trois lettres (Arg, Asn, Asp...). Figure construite par InsightII.

Figure 9: Exemple d'un dictionnaire de rotamères.

Ces informations ont conduit Richards à définir la conformation des chaînes latérales par une combinaison d'angles dièdres préférenciels (appelés rotamères) observés dans les protéines autour d'un seuil de 30degrés (figure 9, Ponder & Richards, 1987). Il a été montré que les rotamères dépendent de la conformation de la chaîne principale (Dunbrack & Karplus, 1993 ; Dunbrack & Cohen, 1997). L'utilisation des rotamères a permis de simplifier la modélisation des chaînes latérales dans les protéines en réduisant l'espace conformationnel à cribler bien qu'il ne soit pas certain que les rotamères couvrent tout l'espace conformationnel (Schrauber et al., 1993). En pratique une technique

d'échantillonnage est couplée à une base de données de rotamère, par exemple : l'algorithme génétique (Tufféry et al., 1991 ; Tufféry et al., 1997), la simulation Monté Carlo (Laughton, 1994a), la minimisation (Dunbrack & Karplus, 1993 ; Chinea et al., 1995 ; Bower et al., 1997) ou encore une technique mixte qui utilise des informations sur l'environement de la chaîne latérale modélisée (Ogata & Umeyana, 1997 ; Ogata & Umeyama, 1998). Il a été montré que la modélisation des chaînes latérales combinant l'utilisation de rotamères avec une optimisation en présence d'un terme d'énergie de solvatation obtenait de meilleurs résultats qu'en phase gazeuse (Cregut et al., 1994). Les résultats montrent que les angles $\chi 1$ sont correctement prédit à 75%, bien que la notion de succès soit difficile à évaluer.

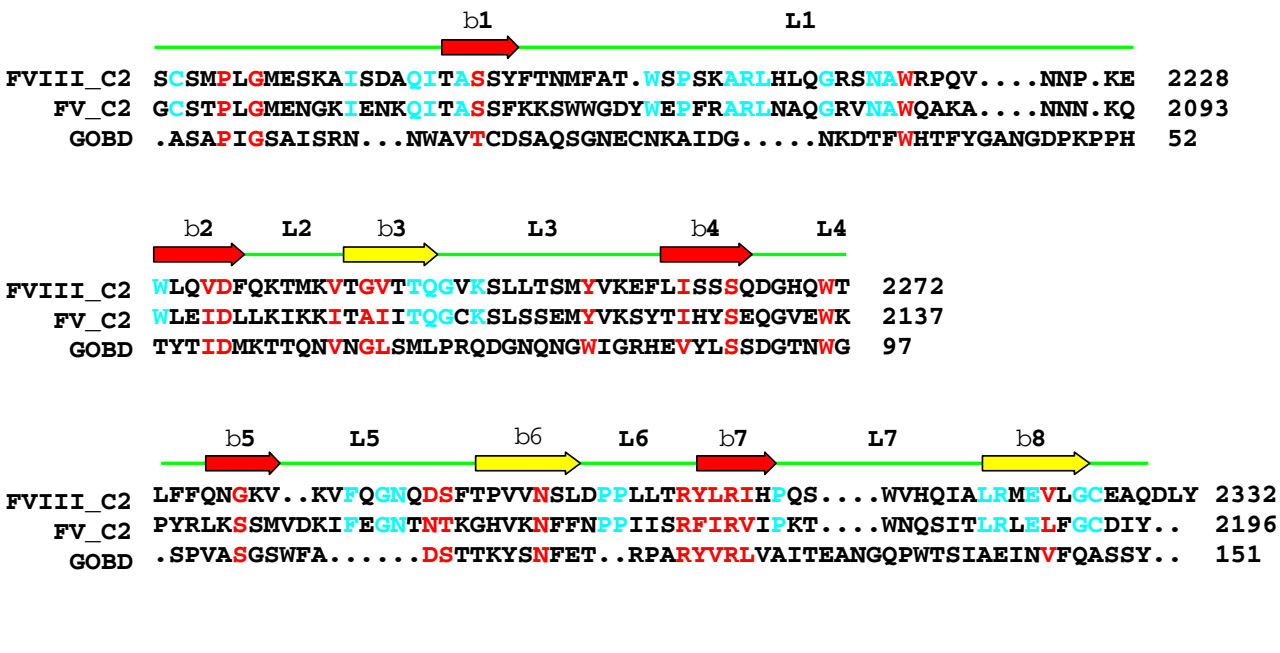
Deux techniques innovatrices déterminent la conformation des chaînes latérales par des techniques de criblages de rotamères. La première utilise un algorithme mathématique (DEE, Dead End Elimination) qui élimine les conformations de rotamères incompatibles entres deux chaînes latérales en contact (Desmet et al., 1992). Des extensions ainsi que des variantes ont été également proposées (Goldstein, 1994 ; Tanimura et al., 1994 ; De Maeyer et al., 1997 ; Lasters et al., 1997). La seconde technique utilise une approche basée sur les champs moyens qui permet de déterminer les rotamères ayant l'énergie la plus favorable au sein d'une matrice conformationnelle comportant tous les rotamères possibles à chaque chaîne latérale (Koehl & Delarue, 1995).

La technique de construction la plus pratiquée cependant est celle proposée par Sutcliffe qui consiste à transférer la conformation des chaînes latérales identiques de la structure parente vers la structure cible, alors que pour les résidus moins homologues la conformation de la structure cible est placée de manière à superposer au mieux celle de la structure parente (Sutcliffe et al., 1987). L'optimisation des chaînes latérales est assuré par ces cycles de minimisation où généralement la position des chaînes latérales conservées sont restraintes dans leurs positions d'origine (Summers & Karplus, 1989 ; Schiffer et al., 1990).

Les chaînes latérales peuvent également être construites par des techniques stochastiques comme la simulation MC à basse température (Shenkin et al., 1996) ou en recuit simulé (Lee & Subbiah, 1991), la simulation MD à haute température (3000K, Laughton, 1994b) avec cependant de moins bons résultats. Une recherche systématique de l'espace conformationnel suivi de minimisations a également été proposé (Eisenmenger et al., 1993). Les résultats sont aussi satisfaisant que les techniques basées sur les rotamères.

5. Construction des insertions/suppressions.

Il est rare de trouver des exemples de modélisation de suppressions. Il est naturellement estimé qu'il est plus facile d'éliminer que d'ajouter. En général les suppressions sont effectuées manuellement grâce à un programme graphique et leur sort dépend largement du type de minimisation employé.
La modélisation des boucles (insertions) est la seconde grande difficulté de cette discipline. Les raisons sont nombreuses : les boucles sont exposées au solvant et par conséquent plus flexibles que le coeur de la protéine, la détermination de la conformation des boucles par RX ou RMN est moins précise que pour le coeur et enfin la majorité des méthodes de prédiction ne prennent pas en compte l'effet du solvant dans la conformation des boucles (figure 10).



Alignement final entre le domaine C2 du FVa et du domaine de liaison de la galactose oxidase. On compte huit "insups" comparés aux 18 présentées par une technique d'alignement automatique standard (cf figure 6). L'alignement multiple avec le facteur de coagulation VIIIa fut aussi très utile. Les "insups" sont localisées uniquement dans les boucles (ligne verte). Les brins de feuillets sont représentés par des flèches.

Figure 10: Insertions et suppressions à modéliser dans le domaine C2 du factor Va.

Une méthodologie consiste à cribler la base de données de structures en fonction de critères géometriques (souvent distances). La difficulté majeure concerne l'insertion de la boucle dans le modèle. En pratique on superpose les extrémités de la boucles sur les extrémités de la protéine cible. Il est donc nécessaire de chercher une boucle ayant une longeur plus grande, en général un résidu à chaque extrémité de la boucle (si on désire

une boucle de cinq résidus, il faut cribler la banque pour une boucle de sept résidus). La complication est qu'il n'y a aucune guarantie que la conformation des deux résidus additionnels soit corrélée à la conformation de la boucle (Tramontano & Lesk, 1992).
Un algorithme génétique a été utilisé pour cribler une base de données de tetrapeptides grâce à un alphabet d'angles dièdres virtuels en Cα (Ring & Cohen, 1994). Une approche par les clusters a également été employée (Fechteler et al., 1995). La combinaison d'une homologie de séquence à la structure secondaire des résidus voisins a été utilisée pour cribler la base de données (Rufino et al., 1997). Les résultats obtenus par ces techniques sont satisfaisants pour les boucles courtes (≤7 résidus) en raison de la saturation de la PDB pour cette taille (Fidelis et al., 1994). Malheureusement, comme nous l'avons indiqué une boucle de sept résidus signifie que l'on ne peut modéliser précisément que des boucles de cinq résidus au plus.

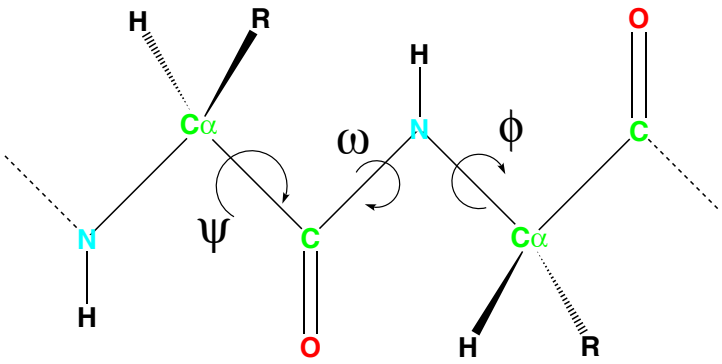
Une autre méthodologie consiste à construire les boucles par une recherche conformationnelle plus ou moins intensive. Les algorithmes de recherche systématique ont l'inconvénient de recourir à la cyclisation de la boucle (Moult & James, 1986 ; Bruccoleri, 1993). Le premier algorithme génère la conformation de la boucle simultanément du côté N et C terminal, la cyclisation est un des paramètres intégrés au criblage des conformations générées (Moult & James, 1986). Le second algorithme génère la conformation de la boucle pour N-3 résidus, les trois derniers résidus étant cyclisés par la technique analytique de Go et Scheraga (1970). La simulation de Monté Carlo en recuit simulé a été employée par plusieurs groupes qui varient en fonction du choix de la conformation de départ de la boucle soit par une conformation aléatoire soit par une conformation étendue (Higo et al., 1992 ; Carlacci & Englander, 1993 ; Collura et al., 1993 ; Carlacci & Englander, 1996). Grâce à une technique de déformation sur sept résidus, il est possible de modéliser une boucle sur le principe de la minimisation globale d'énergie qui consiste à sélectionner de manière cyclique les boucles ayant la plus faible énergie (Dudek & Scheraga, 1990). Les deux méthodes suivantes présentent l'avantage d'éviter l'étape de cyclisation. La première utilise le principe de la relaxation graduée des liaisons covalentes. En traçant une droite entre les deux extrémités de la boucle et en y plaçant les atomes à intervales réguliers, il est possible de construire une boucle par minimisation en augmentant graduellement la taille des liaisons covalentes (Zheng et al., 1993a ; Zheng et al., 1993b ; Rosenbach & Rosenfeld, 1995). La seconde utilise une base de données de dipeptides couvrant les angles dièdres ψ_i,ϕ_{i+1} les plus fréquent au sein de la carte de Ramachandran. En utilisant le recouvrement des angles ψ_i on évite ainsi la cyclisation (Sudarsanam et al., 1995). Ces techniques de recherches ont obtenu de bons résultats et bien qu'il ait été montré que les fonctions énergétiques employées sont appropriées, la technique d'échantillonnage est le point sensible de ces techniques.

6. Optimisation du modèle.

L'optimisation est une étape importante puisqu'elle peut déformer un bon modèle en un modèle médiocre. Généralement, l'optimisation se pratique de paire avec la visualisation et l'analyse géometrique du modèle. Dans un premier temps nous allons décrire les techniques qui permettent d'analyser les modèles en indiquant les régions où la probabilité d'erreur est importante. Ensuite, nous présenterons diverses techniques d'optimisation incluant la minimisation par mécanique moléculaire suivi de conseils pratiques.

a. Analyse d'un modèle moléculaire.

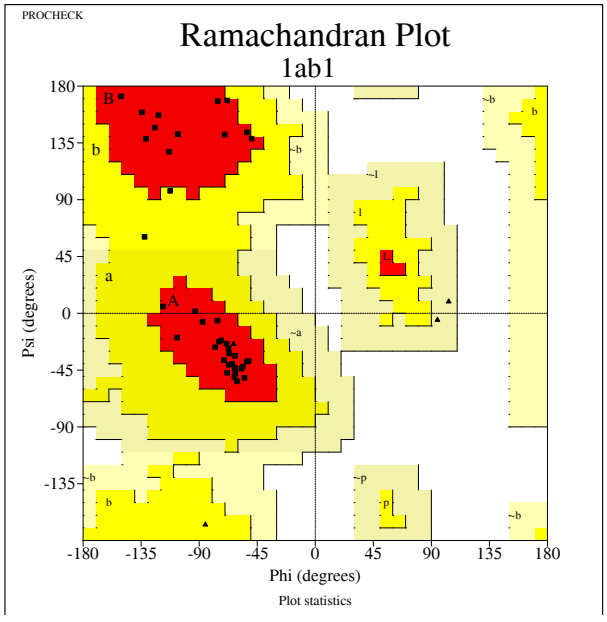
On considère trois niveaux d'analyse : 1) repliement, 2) énergétique, 3) géométrique.
Au niveau du repliement il existe certaines mesures caractéristiques de la conformation des protéines comme le nombre d'Ooi (nombre de Cα présents dans une sphère de rayon r, Nishikawa & Ooi, 1980), le nombre de contacts hydrophobes à l'intérieur des protéines (deux fois plus nombreux chez les protéines natives comparées aux protéines mal-repliées, Bryant & Amzel, 1987), le rayon de giration ($R^2 = 6.9 + 0.967$ N où N est le nombre de résidus, Robson et al., 1987), la surface accessible au solvant ($As = 6.3\ M^{0.73}$ où M est la masse moléculaire pour des protéines ayant une taille comprise entre 4000-35000 Da, Miller et al., 1987), le nombre de résidus exposé à la surface (57% hydrophobe, 24% polaire, 19% chargé, Miller et al., 1987), le nombre de résidus enfoui (58% hydrophobe, 39% polaire, 4% chargé, d'autre part, pour des petites protéines 15% des résidus sont complètement enfoui alors que 32% le sont pour de plus grandes protéines, Miller et al., 1987), ou encore la relation entre l'énergie de solvatation et la taille de la protéine (SFE = 15.30 - 1.13N où N est le nombre de résidus, Chiche et al., 1990).
Au niveau énergétique on distingue deux approches : la première est d'origine statistique et la seconde d'origine thermodynamique. Au niveau statistique, les programmes et les potentiels de "threading" ont été appliqués à l'analyse quantitative de repliement avec une matrice de type 3D-1D, par exemple (Lüthy et al., 1992 ; Kocher et al., 1994) ou encore grâce a une fonction basée sur la distance inter-atomique (Samudrala & Moult, 1998). Les fonctions de potentiels d'énergie sont utilisées très fréquemment pour estimer la stabilité des protéines. Les fonctions qui incluent l'effet de solvant sont de très bons indicateurs puisqu'elles prennent en compte les deux composantes énergétiques les plus défavorables à la stabilité des protéines à savoir l'enfouissement de résidus chargés dans le coeur des protéines et l'exposition de résidus hydrophobe à la surface des protéines (Novotny et al., 1988). Il a été toutefois observé que certaines fonctions avaient des résultats plus ou moins bons que d'autres en fonction du type de protéine analysée (Park et al., 1997). Par conséquent, il est certainement plus prudent d'utiliser plusieurs fonctions énergétiques dans le but d'obtenir des résultats optimaux.



Représentation schématique des angles dièdres de la chaîne principale sur un dipeptide. L'angle ω de la liaison peptidique est majoritairement trans (180 degré) sauf pour certaines prolines (≈0degré). Les angles dièdres préférentiels pour ϕ,ψ sont représentés dans la figure 12. Figure construite par CHEMDRAW.

Figure 11: Représentation schématique des angles dièdres de la chaîne principale.

Au niveau géometrique, il a été montré que la qualité des structures cristallines augmentait avec leur résolution. Les critères géometriques pris en compte sont les angles dièdres de la chaîne principale (ϕ,ψ), des chaînes latérales (χ) et la liaison peptidique (ω), l'angle dièdre ϕ des prolines, la longueur des liaisons hydrogène et la chiralité des ponts disulfures (figure 11, Morris et al., 1992). L'analyse de ces paramètres est grandement facilité par le programme automatique PROCHECK (figure 12, Laskowski et al., 1993).



Exemple de carte de Ramachandran produit par le programme PROCHECK. Un modèle est considéré "bon" lorsque plus de 90% des angles dièdres ϕ,ψ sont localisés dans les régions les plus favorables (rouge). En modélisation, il est plus convenable d'admettre qu'un modèle est bon lorsque le pourcentage d'angles dièdres ϕ,ψ est voisin de celui de la protéine parente.

Figure 12: Angles dièdres préférentiels pour la chaîne primcipale d'une protéine.

Certains paramètres ont été affinés récemment comme l'angle dièdre ω où une plus grande déviation de la moyenne (179.6deg) est observée (±6 au lieu de ±3 auparavant, MacArthur & Thornton, 1996) ou encore par l'étude de résidus déviant des régions favorables de la carte de Ramachandran en raison de tournants de type γ et II', principalement composés de petits acides aminés polaires ou chargés (Gunasekaran et al., 1996). L'utilisation des pseudo-angles dièdres Cα peut se révéler utile (Oldfield94, Kleywegt, 1997). Il est clair que pour la modélisation par homologie les critères géométriques sont les plus précieux bien qu'il soit dangereux dans l'absolue de suivre à tout prix ces critères puisqu'ils peuvent éventuellement fermer la porte à de nouvelles conformations (Eu 3D, 1998).

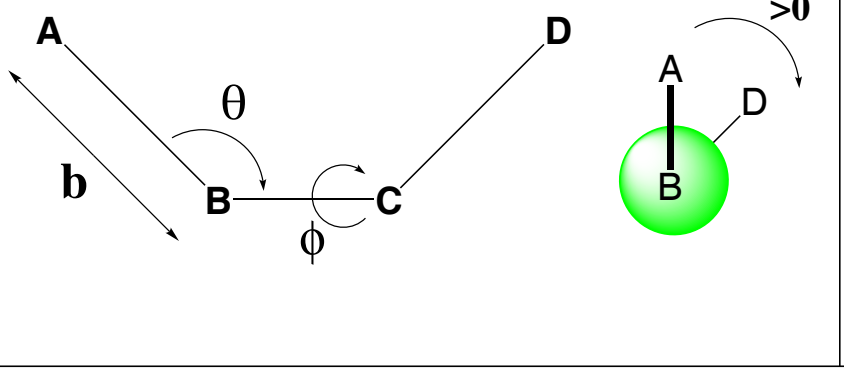
b. Techniques de minimisations.

La minimisation est un procédé mathématique qui essaie de trouver l'énergie minimum globale pour une fonction arbitraire. En mécanique moléculaire cette fonction arbitraire est appelé un champ de potentiels (communément appelé champ de forces). En 1961, Hendrickson montra qu'il était possible de calculer l'énergie d'une molécule par une méthode informatique (Hendrickson, 1961). L'idée évolua pour dévenir un "champ de force de valence" où les fonctions calculent l'énergie d'une molécule en termes de distorsions de la géométrie ajoutés à la somme des interactions entre atomes non liés de manière covalente (Ermer, 1976). La formulation la plus simple, sans termes croisés, est la suivante :

$$E = \sum_{bond} K_b(b - b_o)^2 + \sum_{angl} K_\theta(\theta - \theta_o)^2 + \sum_{dihe} K_\phi(1 + \cos(n\phi + \delta_o)) + \sum_{impr} K_\chi(\chi - \delta_o)^2 + \sum_{i>j} 4\varepsilon_{ij}\left[\left(\frac{r_o}{r_{ij}}\right)^{12} - \left(\frac{r_o}{r_{ij}}\right)^{6}\right] + \sum_{i>j} \frac{q_i q_j}{D r_{ij}}$$

Les termes avec un indice o correspondent aux valeurs géométriques de référence (ou à l'équilibre). Les constantes énergétiques $K_?$ représente l'énergie nécessaire pour déformer les paramètres de leurs positions d'équilibre. Pour l'angle dièdre, n indique la périodicité. Les quatre permiers termes déterminent la déformation de la longueur de la liaison covalente (bond, b), de l'angle de liaison (angl, θ), de l'angle de torsion (dihe, ϕ) et de la planéité (impr, χ, figure 13). Les deux derniers termes concernent les interactions

d'atomes non-liés selon un potentiel de Lennard-Jones (potentiel 12-6) et un potentiel électrostatique (Coulombien où D est la constante diélectrique). Dans le champ de potentiels XPLOR (dérivé de CHARMM,Brooks et al., 1983), par exemple, ε est la profondeur du puits énergétique et r_o la distance minimale entre deux atomes identiques (qui est reliée au rayon de van der Waals par la relation $2R_{vdw} = r_o\ 2^{1/6}$) (Brünger, 1992).



Définition des valeurs géométriques calculées dans un champ de potentiels covalent. Le schéma de droite indique le sens positif de rotation d'un angle dièdre. Figure construite par CHEMDRAW.

Figure 13: Représentation des valeurs géométriques d'un champ de potentiels.

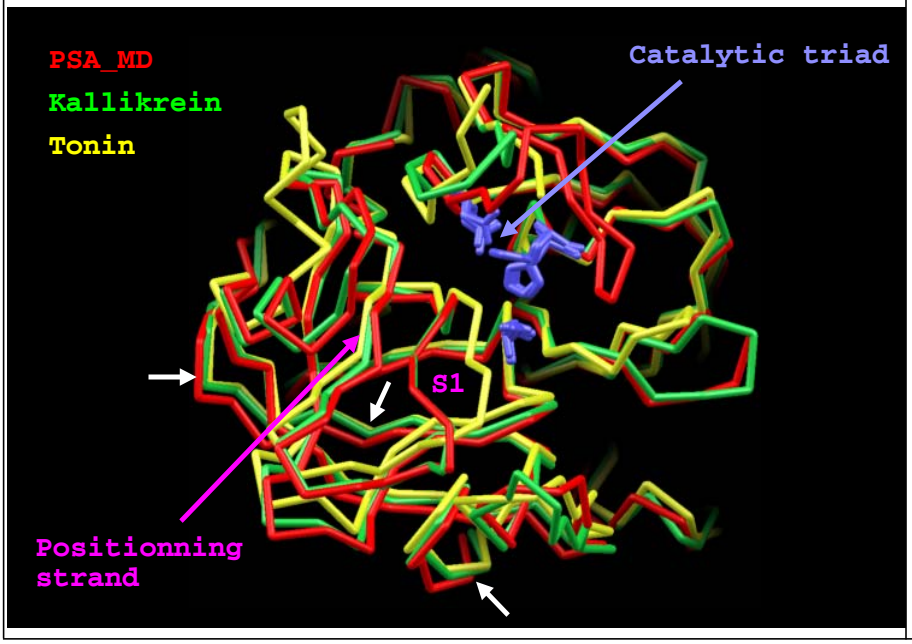
Il exite un grand nombre de champ de potentiels (ECEPP/2, Momany et al., 1975, MM3, Li et al., 1991, TRIPOS5.2, Clark et al., 1989, GROMOS, van Gunsteren et al., 1987, DREIDINGII, Mayo et al., 1990) où certains incluent des termes spécifiques comme des termes croisés (CVFF, Hagler et al., 1974), des termes pour la liaison hydrogène (AMBER, Weiner et al., 1986), des termes précis pour les molécules d'eau (OPLS, Jorgensen & Tirado-Rives, 1988). La première minimisation de modèle moléculaire fut réalisée dans le groupe de Scheraga (Warme et al., 1974).

L'optimisation d'un modèle moléculaire comporte souvent des cycles de minimisation (en phase gazeuseuse), des simulations de Monté Carlo ou de dynamique moléculaire (Garnier, 1990). L'ajout de molécules d'eau explicitement dans le système est encore trop couteux en temps de calcul mais il est possible d'incorporer des termes de solvatation implicite (Abagyan et al., 1994 ; Cregut et al., 1994). Une méthode originale d'optimisation a été proposée par Cachau dans laquelle une carte de densité électronique (basse résolution, ~6Å) fut calculée à partir de la protéine parente ce qui permit de contraindre la dynamique moléculaire. Des factors d'agitation thermique ont pu être calculé grâce à l'analyse de la déviation des atomes durant la simulation ($B = 8\Pi^{2/3}$ <RMSD>, Cachau

et al., 1994).

c. Recommendations.

Le grand danger de l'optimisation par mécanique moléculaire est d'être convaincu que des cycles de minimisations vont améliorer la qualité d'un modèle. En pratique, quelque soit la méthode de minimisation, la structure finale sera toujours très proche de la structure de départ. Une inversion de chaîne latérale ne se produira jamais par exemple (remonté contre le gradient). Les techniques de simulations stochastiques permettent de plus grands mouvements mais leur efficacité est limitée. La raison principale est due à la nature même de la minimisation qui optimise l'énergie totale de la molécule. Par conséquent il est très fréquent d'observer qu'une minimisation intensive sacrifie certains termes énergétiques aux dépens d'autres plus favorables au niveau global. C'est le cas des interactions électrostatiques, qui en phase gazeuse (constante diélectrique=1), domine très nettement la minimisation. Puisqu'en pratique, on n'utilise que rarement un système hydraté, les interactions électrostatiques sont souvent ôtées du champ de potentiels, à l'exception de la minimisation des atomes d'hygrogène dans le but de prendre en compte la nature électrostatique des liaisons hydrogène.
Il apparaît clairement que l'utilisation d'un champ de potentiels incluant tous les atomes (y compris les hydrogènes) est crucial au succès de la minimisation. Il est particulièrement important d'établir le mieux possible l'état de protonation de certains résidus en particulier les histidines.
L'expérience personnelle acquise indique que l'optimisation doit toujours être partielle et contrôlée. Le contrôle s'effectue en imposant des contraintes sur la position des atomes de la chaîne principale par exemple. L'optimisation partielle indique que si la position des atomes de la chaîne principale doit être optimisée (après une modélisation d'"insups" par exemple) alors seule la région concernée doit être optimisée. Il faut être conscient que l'on travaille en trois dimensions et que par conséquent des résidus éloignés dans la séquence peuvent se trouver très proche de l'"insups" et par conséquent doivent être inclus dans l'optimisation.
Le recours à la dynamique moléculaire en phase gaseuze est à éviter autant que possible (figure 14). Dans certaines situations, où la conformation insatisfaisante d'une boucle ne peut être résolue ni manuellement ni par minimisation, le recours à la dynamique moléculaire locale est nécessaire. Dans tous les cas, une visualisation constante de chaque étape de modélisation est impérative. Plusieurs cycles d'affinements sont nécessaires et consistent d'une étape d'optimisation, des étapes de visualisation et des étapes de modifications manuelles de la conformation du modèle.

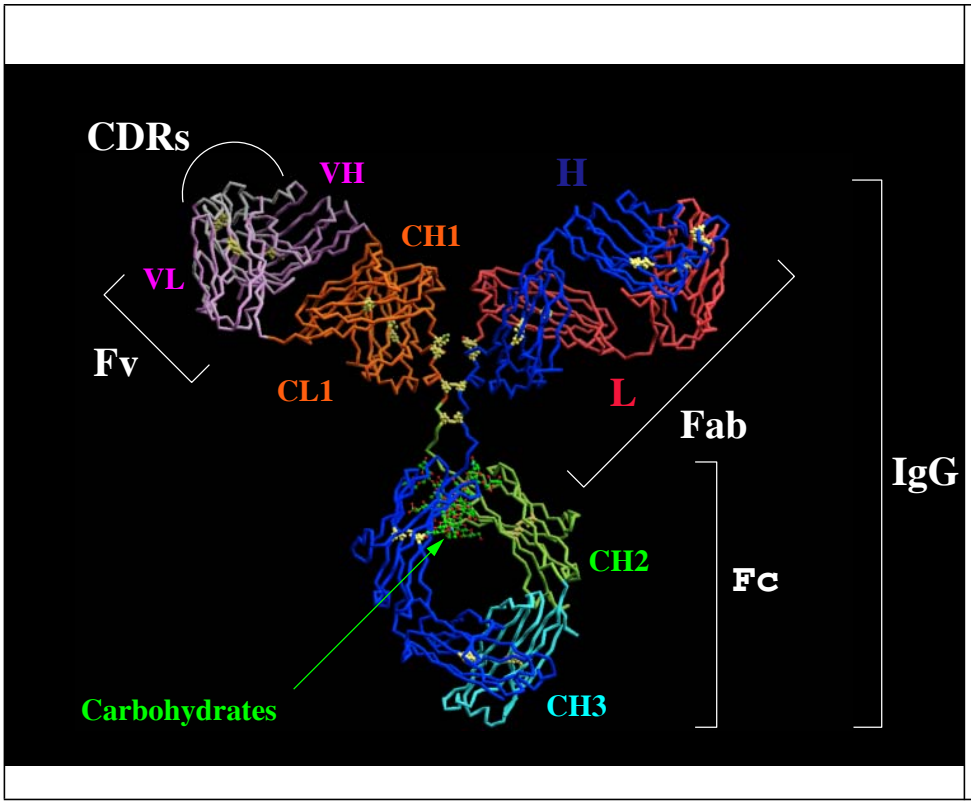


Conséquences d'une sur-optimisation de l'antigène spécifique de la prostate par dynamique moléculaire (Villoutreix et al., 1994). Le brin qui positionne le substrat a dévié suffisamment pour empêcher l'assemblage du peptide substrat. D'autres déviations (par rapport aux molécules parentes) sont indiquées par des flèches blanches. Figure construite par Molscript et Raster3D.

Figure 14: Conséquences structurales d'une sur-optimisation.

7. Modélisation des régions variables d'anticorps.

La modélisation des anticorps est une technique particulière. En effet, bien que chaque séquence d'anticorps (identité > 50%) génère une spécificité differente, la structure des anticorps est particulièrement très conservée même dans les boucles de surfaces (figure 4).
Par anticorps, on sous-entend la partie variable des anticorps sitée à l'extrémité de la molécule (figure 15). Ces régions variables (Fv) sont composées de deux chaînes, une légère (VL) et une lourde (VH), d'environ 110 résidus pour VL et 120 pour VH. Le site de reconnaissance du Fv, qui inclue presque toute les divergences de séquence entre différents anticorps, est composé de six boucles hypervariables (CDRs ou complemantary determining regions, Kabat et al., 1977), trois par chaîne (L1, L2, L3 pour VL, H1, H2, H3 pour VH).



Représentation en tube de la chaîne principale d'une immunoglobuline G. Une des chaînes lourdes (H) est tracée en bleu et une des chaînes légères (L) en rouge. Chaque chaîne est constituée de deux ou quatre domaines d'immunoglobuline (sandwich-β).Le site de reconnaissance de l'antigène est localisé à l'extrémité N terminale de la molécule (CDRs en blanc). La modélisation ne concerne que le domaine qui porte les CDRs, c'est-à-dire le Fv (dessiné en mauve avec les ponts disulfures en jaune). Figure construite par AVS.

Figure 15: Représentation du repliement d'une immunoglobuline.

L'homologie structurale a pour conséquence qu'il n'est plus nécessaire de définir les régions structuralement conservées comme pour les techniques classiques de modélisation. En effet, les séquences sont aisement alignées, si bien qu'un dictionnaire d'alignements existe (figure 16, Kabat et al., 1977). La numérotation des résidus est souvent référée sous le terme Kabat, signifiant que les régions constantes de tous les Fv ont la même numérotation et que les CDRs (boucles) sont définies par des blancs ou des lettres lors d'insertions (par exemple le CDR L1 commence au résidu 26 et fini au résidu 32, incluant ou non des insertions au niveau du résidu 27, tels que 27A, 27B...). La modélisation d'anticorps doit atteindre un niveau de précision unique pour être utile. L'utilité est défini par la possibilité d'assembler l'antigène dans le site de reconnaissance du Fv et de prédire quels résidus seront en contact avec l'antigène. Par conséquent la modélisation correcte des CDRs est cruciale.

Le premier modèle d'anticorps fut le MOPC-315, un anti-haptène, basée sur la chaîne principale de McPC603 (Padlan et al., 1976). Le CDR L1 fut pris du dimère de chaîne légère MCG, les CDRs L2, L3, H1 et H2 proviennent de la molécule parente (McPC603)

```
                                                            L1
                       1                     |----------------|           50
                       1          11      20   24 --------------34         44
1IGC          MOPC21l  NIVMTQSPKS MSMSVGERVT LTCKASENVV ......TYVSWYQQKPEQSP
3MCG             Mcgl  ZSALTQPPS. ASGSLGQSVT ISCTGTS... SDVGGYNYVSWYQQHAGKAP
2RHE             Rhel  ESVLTQPPS. ASGTPGQRVT ISCTGSA... TD.IGSNSVIWYQQVPGKAP
1REI             Reil  DIQMTQSPSS LSASVGDRVT ITCQASQ... ...DIIKYLNWYQQTPGKAP
7FAB            Newml  ASVLTQPPS. VSGAPGQRVT ISCTGSS... SNIGAGHNVKWYQQLPGTAP
2FB4             Koll  ESVLTQPPS. ASGTPGQRVT ISCTGTS... SN.IGSITVNWYQQLPGMAP
8FAB             HILl  ..ELTQPPS. VSVSPGQTAR ITCSANA.LP .....NQYAYWYQQKPGRAP
1IND+         cha2551  ZAVVTQESA. LTTSPGETVT LTCRSST... GAVTTSNYANWVQEKPDHLF
1NGP+           N1G9l  QAVVTQESA. LTTSPGETVT LTCRSST... GAVTTSNYANWVQEKPDHLF
1GIG            HC19l  QAVVTQESA. LTTSPGETVT LTCRSST... GAVTTSNYANWVQEKPDHLF
1MFA+        Se155-4l  QIVVTQESA. LTTSPGETVT LTCRSST... GTVTSGNHANWVQEKPDHLF
1YUH            1YUHl   AVVTQESA. LTTSPGETVT LTCRSST... GAVTTSNYANWVQEKPDRLF
1A6U                    AVVTQESA. LTTSPGETVT LTCRSST... GAVTTSNYANWVQEKPDHLF
1FOR          17-1Al   QIVLTQSPAI MSAFPGEKVT ITCSATS... ....SVNYMHWFQQKPGTSP
3HFL+           Hel5l  DIVLTQSPAI MSASPGEKVT MTCSASS... ....SVNYMYWYQQKSGTSP
1A6T         Mab1-1al  QSVLSQSPAI LSASPGEKVI MTCSPSS... ....SVSYMQWYQQKPGSSP
1AY1             TP7l  DIQMTQSPAI MSASPGEKVT MTCSASS... ....SVSYMYWYQQKPGSSP
1MIM          CHI621l  QIVSTQSPAI MSASPGEKVT MTCSASS... ....SRSYMQWYQQKPGTSP
2FBJ            J539l  EIVLTQSPAI TAASLGEKVT ITCSASS... ....SVSSLHWYQQKSGTSP
1CLO            A5B7l  QTVLSQSPAI LSASPGEKVT MTCRASS... ....SVTYIHWYQQKPGSSP
3HFM           Hel10l  DIVLTQSPAT LSVTPGNSVS LSCRASQ... ...SIGNNLHWYQQKSHESP
1OPG            OPG2l  DELLTQSPAT LSVTPGDSVS LSCRASQ... ...SISNNLHWYQQKSHESP
1MLC+          D44.1l  DIELTQSPAT LSVTPGDSVS LSCRASQ... ...SISNNLHWYQQKSHESP
1GC1                   ELELTQSPAT LSVSPGERAT LSCRASE... ...SVSSDLAWYQQKPGQAP
1JRH              A6l  SVEMTQSPSS FSVSLGDRVT ITCKASE... ...DIYNRLAWYQQKPGNAP
1OSP           184.1l  DIQMSQSSSS FSVSLGDRVT ITCKASE... ...DIYSRLAWYQQKPGNAP
1BBJ           B72.3l  DIQMTQSPAS LSVSVGETVT ITCRASE... ...NIYSNLAWYQQKQGKSP
1FVC+         4D5huml  DIQMTQSPSS LSASVGDRVT ITCRASQ... ...DVNTAVAWYQQKPGKAP
1VGE           TR1.9l  ELVMTQSPSS LSASVGDRVN IACRASQ... ...GISSALAWYQQKPGKAP
1FIG             1F7l  ENVLTQSPAI MSASPGEKVT MACRASS... ..SVSSTYLHWYQQKSGASP
1A2Y            D1.3l  DIVMTQSQKF MSTSVGDRVS ITCKASQ... ...NVGTAVAWYQQKPGQSP
6FAB            3671l  DIQMTQIPSS LSASLGDRVS ISCRASQ... ...DINNFLNWYQQKPDGTI
1KNO+         CNJ206l  QIQMTQSPSS LSASLGERVS LTCRASQ... ...EISGYLSWLQQKPDGTI
1GAF+           48G7l  DIQMTQSPSS LSASLGERVS LTCRASQ... ...EINGYLGWLQQKPDGTI
1GHF          GH1002l  DIQMTQTTSS LSASLGDRVT ISCRESQ... ...DISNSLNWYQQKPDGTV
1DVF            E5.2l  DIQLTQSPSS LSASLGDRVT ISCRASQ... ...DISNYLNWYQQKPDGTV
1IKF             R45l  DIQMTQTTSS LSASLGDRVT ISCRASQ... ...DISTYLNWYQQKPDGTV
1JHL          D11.15l  DIELTQSPSY LVASPGETIT INCRASK... ...SISKSLAWYQEKPGKTN
1EAP            17E8l  DIELTQSPSS LSASLGGKVT ITCKASQ... ...DIKKYIGWYQHKPGKGP
1NCA+           NC41l  DIVMTQSPKF MSTSVGDRVT ITCKASQ... ...DVSTAVVWYQQKPGQSP
12E8                   DIVMTQSQKF MSTSVGDRVS ITCKASQ... ...NVGTAVAWYQQKPGQSP
1AXS                   ELVLTQSPSS MYASLGERVT ITCKASQ... ...DINSYLNWFQQKPGKSP
1FGV          H52huml  DIQMTQSPSS LSASVGDRVT ITCRASQ... ...DINNYLNWYQQKPGKAP
1IAI         409.5.3l  DIQLTQSPAF MAASPGEKVT ITCSVSS... ..SISSSNLHWYQQKSETSP
1GGC+           50.1l  DIVLTQSPGS LAVSLGQRAT ISCRASESVD ..DDGNSFLHWYQQKPGQPP
1AFV            25.3l  DIVLTQSPAS LAVSLGQRAT ISCRASESVD ..NYGISFMNWFQQKPGQPP
1ACY            50.9l  DIVMTQSPAS LVVSLGQRAT ISCRASESVD ..SYGKSFMHWYQQKPGQPP
1IBG           40-50l  .IVLTQSPAS LAVSLGQRAT ISCRASKSVS TS..GYSHIHWYQQKPGQPP
1NMB+           NC10l  DIVLTQSPSS LAVSLGQRAT ISCRASQSVS TS..SFRYMHWYQQKPGQPP
1YEC+           D2.3l  DIVMTQSPLT LSVTIGQPAS ISCKSSQSLL YS.NGKTYLNWLLQRPGQSP
1HYX(DMV)        6D9l  ELVMTQTPLS LPVSLGDQAS ISCRSSQTIV HS.NGDTYLDWFLQKPGQSP
1RMF            R6.5l  DVVMTQSPLS LPVSLGDQAS ISCRSSQSLV HS.NGNNYLHWYLQKSGQAP
1FPT              C3l  DVVMTQTPLS LPVSLGDQAS ISCSSSQSLV HS.NGKTYLHWYLQKPGQSP
1PLG            1PLGl  DVVMTQTPLS LPVSLGDQAS ISCRSSQSLV HS.NGNTYLYWYLQKPGQSP
1TET            TE33l  DVLMTQTPLS LPVSLGDQAS ISCKSSQSIV HS.GGNTYFEWYLQKPGQSP
1IGI           26-10l  DVVMTQTPLS LPVSLGDQAS ISCRSSQSLV HS.NGNTYLNWYLQKAGQSP
....            Bv33l  DVVMTQTPLS LPVSLGDQAS ISCRSSQSLV HS.NGNTYLHWYLQKPGQSP
1FLR           4-4-20l DVVMTQTPLS LPVSLGDQAS ISCRSSQSLV HS.QGNTYLRWYLQKPGQSP
1DBB+            Db3l  DVVMTQIPLS LPVNLGDQAS ISCRSSQSLI HS.NGNTYLHWYLQKPGQSP
1MRD+         JEL103l  DVVMTQTPLS LPVSLGDQAS ISCRSSQSLV HS.NGNTYLHWYLQKPGQSP
1NBV+        Bv04-01l  DVVMTQTPLS LPVSLGDQAS ISCRSSQSLV HS.NGNTYLHWYLQKPGQSP
1CFV            4155l  DIELTQSPPS LPVSLGDQVS ISCRSSQSLV SN.NRRNYLHWYLQKPGQSP
2JEL           JEL42l  DVLMTQTPLS LPVSLGDQAS ISCRSSQSIV HG.NGNTYLEWYLQKPGQSP
2H1P                   DVVMTQTPLS LPVSLGDPAS ISCRSSQSLV HS.NGNTYLHWYLQKPGQSP
1NLD            1583l  DVVMTQTPLT LSVTIGQPAS ISCKSSQSLL DS.DGKTYLNWLLQRPGQSP
2CGR            1CGSl  ELVMTQSPLS LPVSLGDQAS ISCRPSQSLV HS.NGNTYLHWYLQKPGQSP
1A4J                   ELVMTQTPLS LPVSLGDQAS ISCRSSQSLV HS.NGNTYLHWYLQKPGQSP
1AXT           33F12l  ELVMTQTPLS LPVSLGDQAS ISCRSSQSLV HS.YGNTFLNWYLQKSGQSP
1CLY+           CBR96l ..LMTQIPVS LPVSLGDQAS ISCRSSQIIV HN.NGNTYLEWYLQKPGQSP
1CLZ            MBR96l DVLMTQIPVS LPVSLGDQAS ISCRSSQIIV HN.NGNTYLEWYLQKPGQSP
1IGF+           B13i2l DVLMTQTPLS LPVSLGDQAS ISCRSNQTIL LS.DGDTYLEWYLQKPGQSP
1KEM+            28B4l DVLMTQTPLS LPVSLGDQAS ISCRFSQSIV HS.NGNTYLEWYLQKSGQSP
1HIL             17/9l DIVMTQSPSS LTVTAGEKVT MSCTSSQSLF NSGKQKNYLTWYQQKPGQPP
1A3R+             8F5l DIVMTQSPSS LTVTTGEKVT MTCKSSQSLL NSRTQKNYLTWYQQKPGQSP
1MCP+            1mcpl DIVMTQSPSS LSVSAGERVT MSCKSSQSLL NSGNQKNFLAWYQQKPGQPP
```

Exemple d'alignement multiple de séquences d'anticorps indiquant leur très grande homologie dans le repliement des immunoglobulines.
Les variations se localisent principalement dans les boucles chargées de reconnaitre l'antigène, les CDRs (içi, seulement le CDR L1 est représenté).
La première colonne indique le code du fichier PDB associé à la séquence. Le signe + indique qu'il existe plusieurs structures pour cette séquence. La seconde colonne indique le nom commun de l'anticorps s'il est connu.

Figure 16: Variation de séquences autour du CDR L1 d'anticorps.

bien que des ajustements furent nécessaire pour H1 (insertion) et pour H2 (suppression de trois résidus). Le CDR H3 provient d'un mélange entre McPC603 et le dimère de VL NEW. Le modèle mécanique a permis de proposer un mode de liaison de l'haptène DNP (Padlan et al., 1976). A partir de ce modèle on peut conclure que la modélisation des CDRs est triviale puisque qu'ils proviennent tous d'autres anticorps et par conséquent, ils ne requièrent ni cyclisation, ni simulation. Toutefois, si l'orientation des CDRs est grandement facilitée, le positionnement des chaînes latérales est la clé du succès. Malheureusement, même à l'heure actuelle avec plus d'une centaine de structures d'anticorps, le positionnement des chaînes latérales reste problématique.
Le second modèle fut le Fv J539, un anti-oligosaccharide construit à partir de McPC603 (Feldmann et al., 1981) affiné ultérieurement (Mainhart et al., 1984). D'autres modèles incluent des Fv anti-lysozyme (Gloop1-5) contruits à partir de NEW (de la Paz et al., 1986) ou encore HyHEL10 construit à partir de McPC603 (Smith-Gill et al., 1987), des anti-oligosaccharides (W3129, 19.1.2) construits à partir de McPC603 et J539 (Padlan & Kabat, 1988).

En 1986, Chothia démontra que la conformation des CDRs est relativement conservée et que l'on peut prédire la conformation de cinq d'entre eux (à l'exception du CDR H3) simplement par l'identification de certains résidus localisés à des positions clés dans la structure (Chothia & Lesk, 1986). Cette technique a permis de modéliser la conformation de l'anticorps D1.3 (à partir de REI et de KOL) avant que sa structure soit expérimentalement résolue. Le RMSD des CDRs fut de l'ordre de 0.5 à 2.07 Å (Chothia & Lesk, 1986). Ces conformations standards ont été nommées : conformations canoniques des CDRs (Chothia & Lesk, 1987). Pour le CDR L1 avec une chaîne légère λ, un tournant de type I est observé entre les résidus 26 et 29 alors que les résidus de 27 à 30B forment une hélice irrégulière; de plus les résidus 25, 30, 33, 71 participent à la conformation de ce CDR. Pour la chaîne κ du CDR L1, les résidus entre 26 et 28 sont en conformation étendue et les résidus 29-32 forment un tournant déformé de type II (résidus 2, 25, 33, 71 sont également important). La conformation du CDR L2 est très conservée sous la forme d'un tournant à trois résidus. Le CDR L3 forme une liaison hydrogène entre les résidus 92 et 95 de la chaîne λ, alors que les résidus 93-96 sont en conformation étendue pour la chaîne κ en raison de la présence d'une proline, à la position 95, qui possède une liaison peptidique cis (si une proline apparaît à la position 94 alors un tournant serré est observé). Le CDR H1 est caractérisé par l'enfouissement du résidu 29 contre les résidus 34, 72, 77, alors que le résidu 27 est partiellement enfoui contre le résidu 94. Le CDR H2 varie beaucoup en taille et forme : un tournant à trois résidus, à quatre résidus (sans liaison hydrogène) ou un tournant serré à six résidus. Ces définitions ont été étendues plus tard où le nombre des structures canoniques s'élève à 4 pour L1, 1 pour L2, 3 pour L3, 2 pour H1 et 4 pour H2 (Chothia et al., 1989 ; Martin & Thornton, 1996). En terme de modélisation, l'utilisation des structures canoniques est simple ; une fois identifié

l'appartenance du CDR à une classe canonique, il suffit de prendre le prototype de cette classe comme CDR.

Pourquoi est-il aussi difficile de modéliser les CDRs d'anticorps alors que cinq d'entre eux ont des conformations conservées et prédictibles ?
La réponse est simple : CDR H3. Le CDR H3 est le plus variable des CDRs à la fois en taille (de 4 à 17 résidus) et en conformation. C'est pourquoi aucune structure canonique n'a pu être identifiée. Mais la raison principale provient du fait de la position centrale qu'occupe le CDR H3 par rapport au site de reconnaissance. Par conséquent, presque tous les CDRs forment des contacts avec H3 et une erreur dans H3 sera propagée rapidement chez les autres CDRs. Une autre difficulté provient du fait qu'en raison de la position centrale qu'occupe le CDR H3, il affecte très sensiblement l'orientation entre VL et VH. Il a été montré que la flexibilité de l'interface VL-VH était un critère important lors des changements conformationnels obervés lorsque l'antigène se lie à l'anticorps (Pellequer et al, 1996).

Pour toutes ces raisons, la modélisation des sites de reconnaissance d'anticorps est pratiquée de manière artisanale, où chaque anticorps à sa propre caractéristique et sa propre particularité de séquence, empêchant le développement de programmes automatiques efficaces de prédictions bien que certains existent (Viswanathan et al., 1995). Plusieurs programmes d'échantillonnage d'espace conformationnel ont été utilisés pour modéliser les CDRs (Fine et al., 1986 ; Moult & James, 1986 ; Shenkin et al., 1987 ; Bruccoleri et al., 1988 ; Mas et al., 1992). Une technique mixte de prédiction a été proposée par Martin qui combine l'utilisation du programme d'échantillonnage CONGEN à une recherche dans les bases de données structurales (Martin et al., 1989 ; Martin et al., 1991). Un concept de modélisation a été introduit par Roberts où la modélisation d'anticorps est affinée grâce à une base de données structurales d'anticorps similaires (Roberts et al., 1994). Cette procédure a été appliquée pour modéliser plusieurs anticorps (le Fv catalytique 43C9, Roberts et al., 1994, des anti-phosphocholine T15 et D16, Chen et al., 1995 ; Brown et al., 1996, et un anti-diuron, Bell et al., 1995).
En raison du grand nombre de structures d'anticorps connus, il est désormais possible d'analyser statistiquement la conformation du CDR H3 afin d'en déduire des règles de modélisation (Shirai et al., 1996 ; Morea et al., 1998 ; Oliva et al., 1998). Cependant, la présence d'exceptions empêche la généralisation de ces règles. Le développement d'une technique basée sur les réseaux neuronaux ne fournit guère de solutions satisfaisantes (Reczko et al., 1995).

8. Estimation en aveugle de la modélisation par homologie.

John Moult organisa en 1994 à Asilomar le premier meeting d'analyse du succès des

techniques de modélisation (CASP: Critical Assessment of Structure Prediction) en fournissant la séquence protéique de molécules cibles pour lesquelles des structures cristallines furent en cours de résolution. Au premier meeting, 13 laboratoires ont présenté 43 modèles pour les huit molécules cibles (figure 17). Un deuxième meeting s'est tenu à Asilomar en 1996 où 19 laboratoires ont présentré 48 modèles pour les neuf molécules cibles (figure 18). Le troisième meeting vient de s'achever en 1998. Les résultats des meetings sont accessibles sur le web du centre CARB (http://iris4.carb.nist.gov/casp).
Les résultats des deux premiers meetings sont représentés dans les figures 17 et 18 et sont publiés dans deux éditions spéciales du journal *Proteins* (vol 23, numéro 3 en 1995, et Suppl. 1 en 1997). Sans surprise, les résultats indiquent que plus l'homologie est grande meilleur est le modèle. La gamme d'erreur, très large pour des protéines ayant des homologies entre 30 et 50%, fut inattendue. Une autre surprise est qu'il n'y a pas vraiment de techniques gagnantes, certaines s'appliquent mieux à certaines protéines que d'autres. (Mosimann et al., 1995 ; Martin et al., 1997).
Trois conclusions émergent de ces meetings : presque toutes les erreurs des modèles sont dues à un mauvais alignement entre les séquences cibles et parentes ; les plus grandes erreurs sont localisées dans les boucles ; plusieurs structures derivent nettement de la structure parente indiquant une sur-optimisation.
En quelques lignes, voici les recommendations d'un des participants aux meetings (Abagyan et al., 1997):
1) Prendre la structure parente la plus homologue possible.
2) Modifier l'alignement pour minimiser les insertions dans le coeur et la taille des "insups".
3) Analyser la structure de la protéine parente et déterminer quelles sont les régions de la chaîne principale qui peuvent dévier de leurs positions initiales.
4) Placer les chaînes latérales non homologues dans la configuration la plus probable statistiquement (premier rotamère).
5) S'arrêter là.

Code	Cible	% id	Taille	RMSD (Cα)	Parent
NM23	NDP kinase	77	148	0.53-1.18	1NDL
E5.2	Immunoglobulin domain	76	228	1.35-1.47	1FAI
Hpr	Phosphocarrier protein	42	88	0.79-1.27	2HPR
CRABPI	Lipocalin	41	136	2.01-3.72	2HMB
HFD	Ferredoxin	40	128	9.94	1FXA
EDN	RNase	35	134	2.85-5.22	6RSA
P450	Cytochrome	22	403	4.25-7.40	1CPT

Figure 17: Molécules cibles et résultats du meeting CASP1.

Code	Cible	% id	Taille	RMSD (Cα)	Parent
T0001	Dihydrofolate reductase	34	162	1.89-3.75	7DFR
T0002	Threonine deaminase	22	514, 186	6.25	1WSY_B
T0003	Glucose permease	44	154	1.34-3.06	1GPR
T0004	Polyribonucleotide nucleotidyltransferase	24	84	7.38	1CSP
T0009	Stellacyanin	26	108	2.28-2.81	2CBP
T0017	Gluthiaone transferase	85	217	0.41-2.69	2GST_A
T0024$^{(1)}$	UBC9	37		2.39-3.48	1AAK
T0024$^{(2)}$	UBC9	37			1AAK
T0027$^{(1)}$	Pectate lyase A	20	319	18.68	2PEC
T0027$^{(2)}$	Pectate lyase A	20			2PEC
T0028	Endoglucanase I	47	359	3.37-5.26	1CEL_A

Figure 18: Molécules cibles et résultats du meeting CASP2.

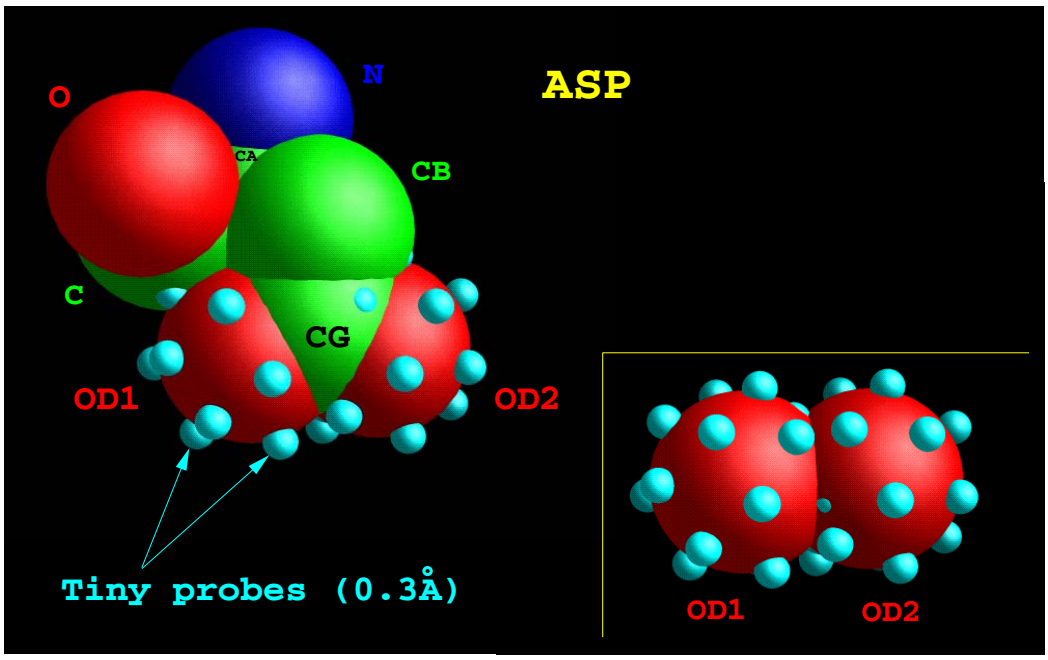
CONTRIBUTIONS PERSONNELLES

I. Développement de techniques d’analyses et d’évaluations énergétiques des protéines.

1. Outils d’analyse de la conformation des protéines.

a. Méthode originale pour calculer la compacité des atomes**.**

La reconnaissance d'un antigène par un anticorps requiert un contact intime entre les chaînes latérales de chacun des partenaires. La comparaison entre les structures cristallographiques d'anticorps dans leur forme libre ou complexée à leur antigène, révèle des différences structurales dans la conformation de l'anticorps. Ces changements conformationnels sont variés en nature et en amplitude. En effet, on trouve des rotations de chaînes latérales de 1 à 180 degrés, des mouvements de boucles hypervariables, ou bien des changements dans l'orientation relative des domaines chaîne légère (VL) et chaîne lourde (VH).
Dans le but d'étudier de manière quantitative ces changements conformationnels (Pellequer et al. 1996) , nous avons développé un programme qui permet de calculer la compacité des atomes dans une macromolécule. L'algorithme est innovateur puisqu'il permet de détecter des changements conformationnels relatifs d'atomes ce qui n'est pas possible par la méthode classique du calcul de la déviation des écarts quadratiques moyens (RMSD) entre des atomes ayant une conformation A vis-à-vis d'atomes ayant une conformation B. Notre méthode permet, non seulement d'identifier quels atomes se rapprochent ou s'éloignent les uns des autres, mais également de calculer la surface de contact gagnée ou perdue, lors du changement conformationnel. L'algorithme utilise une sonde sphérique, de rayon déterminé par l'utilisateur, qui "roule" à la surface de chaque atome de la molécule (figure 19). Lorsque notre sonde entre en contact avec un atome voisin (non covalemment lié), on assigne un point de contact à cet atome (on conserve l'identité de l'atome contacté). A chaque point conservé on détermine une surface de contact grâce à un coefficient multiplicateur qui dépend directement de la densité de points sélectionnée par l'utilisateur. L'avantage de cette méthode est qu'elle ne requiert pas de connaissance a priori sur les acides aminés telle que le volume exact d'une chaîne latérale.



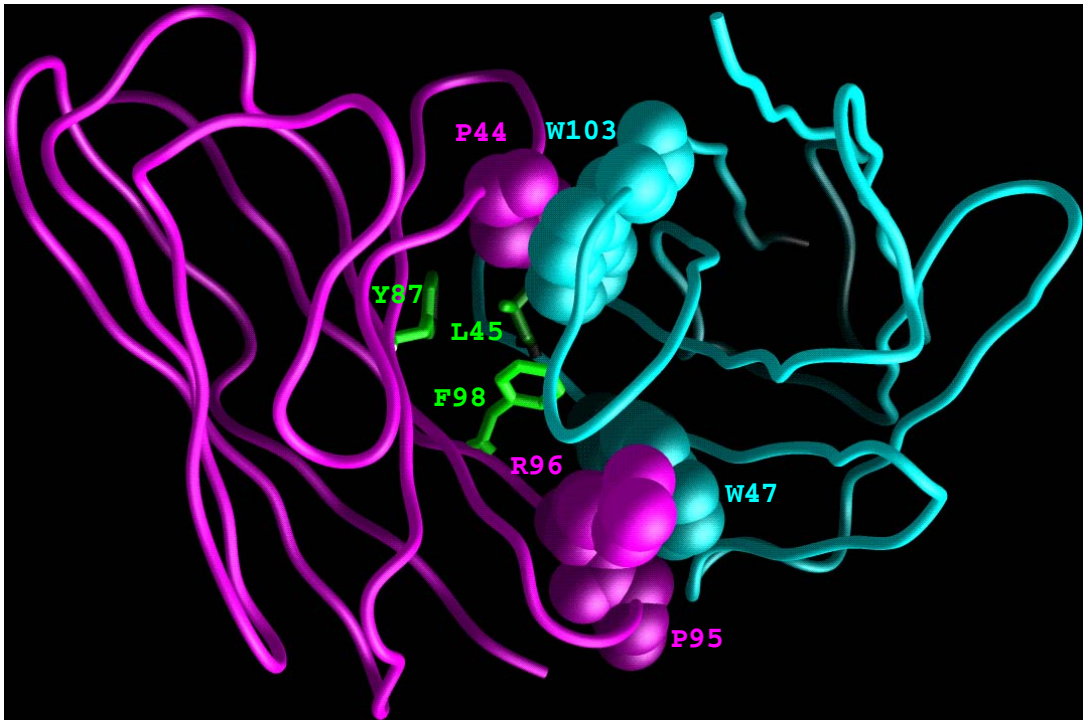
Les positions visitées par la sonde sont représentées en cyan. Seules les sondes qui contactent au moins deux atomes non liés seront conservées ultérieurement. Pour chaque position conservée une surface de contact est sauve-gardée. Figure construite par InsightII.

Figure 19: Principe de fonctionnement du programme Tiny.

b. Analyse des changements conformationnels à l'interface VL-VH.

Notre méthode permettant de calculer la compacité des atomes nous a permis d'étudier les changements conformationnels se produisant à l'interface des chaînes légères et lourdes (VL-VH) d'anticorps lors de la reconnaissance antigénique. Nous avons analysé 12 complexes d'anticorps dont on connaît à la fois la conformation liée et non liée à l'antigène. Les résultats indiquent qu'il est possible de caractériser les changements conformationnels de l'interface VL-VH sous trois formes : une interface moins compacte, une interface plus compacte et une interface neutre. Lorsque des antigènes de plus de 24 atomes, mais non protéiques, se lient aux anticorps, ils provoquent une ouverture de l'interface (moins compacte) alors que des antigènes de moins de 24 atomes provoquent une fermeture de l'interface (plus compacte). Lorsque des proteines se lient aux anticorps, elles ne provoquent pratiquement pas de changement de l'interface VL-VH. Cette corrélation entre la taille de l'antigène et le type de changement conformationnel observé chez l'anticorps suggère que l'interface joue un rôle important

dans le processus de la reconnaissance moléculaire. Puisque que notre outil d'analyse nous permet d'obtenir des changements conformationnels sous forme de surface de contacts, il est possible de traduire ces valeurs en terme d'énergie d'association. Les valeurs indiquent qu'il est possible de perdre jusqu'à 3.4 kcal/mol entre VL et VH, soit une diminition de l'affinité de l'ordre de 300 fois. Par conséquence on postule que des mutations localisées à l'interface VL-VH (sans contact direct avec l'antigène) peuvent moduler l'affinité d'un anticorps avec son antigène (figure 20). Il semble judicieux que des techniques de criblage d'anticorps (banques de phages) incluent certains résidus localisés à l'interface VL-VH dans le but d'accroître la diversité de leurs banques.



Résumé des changements majeurs qui se produisent à l'interface VL (magenta) et VH (cyan). Les paires de résidus L44-H103, L95-H47, L96-H47, L87-H45, L98-H45 montrent les plus grands changements parmi nos 12 anticorps analysés. Figure construite par InsightII.

Figure 20: Principaux changements au niveau de l'interface VL-VH d'anticorps.

2. Développement de fonctions énergétiques calculant l'énergie conformationnelle (Article no1, annexe I).

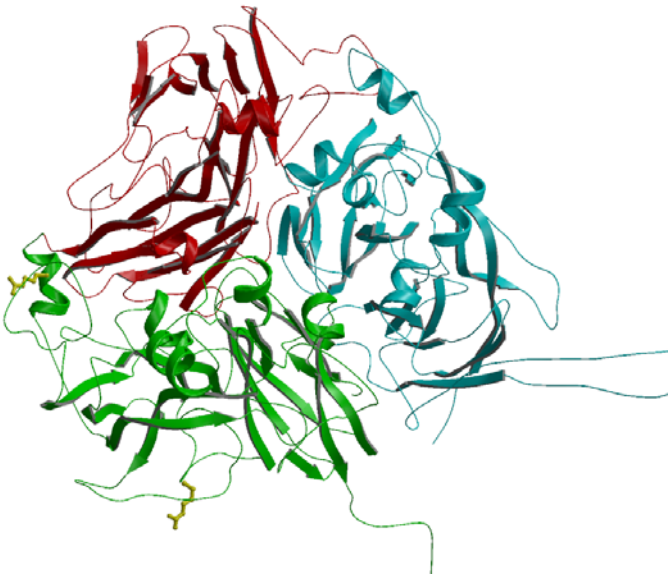
Pellequer J.L. & Chen S.w.W. (1997).
Does conformational free energy distinguish loop conformations in proteins? Biophys. J. **73**, 2359-2375.

Le but de cette étude était de pouvoir sélectionner les boucles d'une base de données susceptibles d'avoir une conformation similaire à celles de boucles hypervariables d'anticorps (CDRs) une fois introduites dans un anticorps. On a simulé une modélisation de boucles hypervariables en remplaçant celles d'un anticorps de structure connue par les boucles de notre base de données. L'anticorps sélectionné, également appelé structure native par opposition à une structure modèle, est le R45-45-11 dont la structure tridimensionnelle a été résolue dans le laboratoire d'Immunochimie de l'IBMC par D. Altschuh.
La différence d'énergie conformationnelle, entre la structure modèle et la structure native, en solution est égale à l'énergie conformationnelle en phase gazeuse (vide) plus la différence d'énergie de solvatation. Le solvant est représenté par un modèle continu, i.e. constante diélectrique de 80. L'énergie de solvatation est décomposée de la manière suivante : une composante électrostatique et une composante non polaire. La contribution électrostatique est obtenue par la résolution de l'équation de Poisson-Boltzmann dans l'espace tridimensionnel par une méthode numérique (FDPB). La contribution non polaire (ou hydrophobe) est obtenue par le produit d'un coefficient d'énergie de transfert d'une phase gaseuze vers un solvant aqueux (terme $\gamma = 0.005$ kcal/mol/Å^2) à la surface totale accessible au solvant. Les charges atomiques de l'anticorps sont "immergées" dans un milieu continu de faible constante diélectrique (ε=2) délimité par la surface moléculaire de l'anticorps.
Nos résultats indiquent qu'il est possible de discriminer systématiquement par le calcul de l'énergie conformationnelle en phase gazeuse les boucles ayant les plus faibles RMSD comparées aux boucles "natives" de l'anticorps. On a montré que l'effet de solvatation est nécessaire dans certains cas pour discriminer les boucles ayant une conformation incorrecte et une très faible énergie en phase gazeuse (e.g. CDRL1). Nous avons également observé que l'ajout de l'énergie de solvatation peut être néfaste (e.g. CDRH3) où des boucles ayant des enthalpies libres conformationnelles élevées en phase gazeuse apparaissent les plus stables en solution, alors qu'elles n'ont aucun caractère commun avec la boucle CDRH3 native. L'analyse quantitative de ce dernier phénomène est en cours d'étude.

II. Modèles moléculaires à moyenne résolution : Etudes de mécanismes fonctionnels.

1. Modélisation des domaines A du facteur de coagulation Va.

Ce projet consiste à étudier au niveau moléculaire le rôle du cofacteur Va dans le complexe de la protrombinase qui comprend, outre le cofacteur Va (FVa), le facteur de coagulation Xa (FXa), la protrombine, des atomes de calcium et la présence de lipides négativement chargés. Le complexe protrombinase est la dernière étape de la cascade d'événements générant la trombine qui à son tour va cliver le fibrinogène dans le but de créer un caillot sanguin. Le FVa a cinq domaines dénomés A1 et A2 (chaîne lourde), A3, C1 et C2 (chaîne légere). Les domaines homologues "A" (1022 résidus) ont été modélisés à partir de la ceruloplasmine (39% de résidus identiques), une protéine transportant des atomes de cuivre dont on connaît la structure tridimensionnelle : un trimère symétrique de sous-unités non identiques (Pellequer et al, 1998). Chaque domaine "A" est composé de deux tonneaux β de type plastocyanine. Lorsque les trois domaines "A" sont assemblés, ils forment un cône tronqué avec une large base plane (figure 21). Le complexe protrombinase est régulé par protéolyse limitée du FVa, à deux endroits (Arg 306 et Arg 506), par la protéine C activée (APC). Ces deux résidus sont distants d'environ 40Å sur le même coté du FVa. On sait que le clivage à la position R506 est plus rapide que celui de R306, ce qui est tout à fait cohérent avec notre modèle puisque R506 est complètement exposée au solvant alors que R306 est partiellement enfouie. Parmi les huit atomes de cuivre présents dans la ceruloplasmine, quatre sont définitivement absents dans le FVa. Il est intéressant d'observer que parmi les sites d'atomes de cuivre restant, tous se trouvent localisés à l'interface entre la chaîne lourde (A1) et la chaîne légère (A3). Ceci peut s'avérer important lors de la dissociation du FVa après protéolyse complète. Ce modèle devrait permettre de mieux comprendre la régulation du complexe protrombinase. De plus, il devrait nous aider à orienter les autres partenaires du complexe (FXa, protrombine, APC, membrane lipidique) grâce au positionnement de l'APC sachant que cette dernière est liée à la membrane.



Vue du dessus du cône tronqué des domaines “A” du FVa. Les domaines A1 (rouge) et A2 (cyan) forment la chaîne lourde alors que A3 forme une partie de la chaîne légère. Les arginines 306 et 506, impliquées dans le processus d’inactivation du Fva, sont dessinées en jaune. Ce modèle comporte 16318 atomes. Figure construite par Molscript et Raster3D.

Figure 21: Modèle des domaines A du facteur Va.

2. Modélisation des domaines C du facteur de coagulation Va (Article no2, annexe I).

Pellequer J.L., Gale A.J., Griffin J.H. & Getzoff E.D. (1998).
Homology models of the C domains of blood coagulation Factors V and VIII: A proposed membrane binding mode for FV and FVIII C2 domains. Blood Cells, Mol. Diseases. **24**, 448-461.

Récemment nous avons identifié, grâce à une technique de “threading”, un prototype putatif de la structure tridimensionnelle des domaines C1 et C2 de FVa. Il s’agit du domaine de liaison de l’enzyme galactose oxidase (GOBD) qui contient 151 residus. L’homologie de séquence entre le domaine C2 et le GOBD étant plus grande que celle entre le domaine C1 et GOBD, nous avons décidé de modéliser le domaine C2 dans un premier temps. Sachant que le domaine C1 possède une homologie de séquence de l’ordre de 50% vis-à-vis du domaine C2, le domaine C1 sera modélisé à partir du domaine C2. Un alignement automatique de séquence entre le domaine C2 et GOBD indique la présence de 18 “GAPS” (figure). Cependant, en utilisant la structure secondaire connue de GOBD et en n’autorisant la présence d’insertions/délétions uniquement dans les boucles, le nombre de “GAPS” chute à 8 (figure). Le repliement des domaines C est un “β-sandwich” où un feuillet à cinq brins fait face à un feuillet à trois brins. Le contenu en boucle est de 66% ce qui permet d’expliquer la difficulté rencontrée lors de la recherche de protéines homologues dans la base de données PDB. Le domaine forme schématiquement un cylindre d’une longueur de 40Å et d’un diametre de 30Å. Les domaines C1 et C2 sont covalemment attachés tête-beche pour former un cylindre d’environ 80Å de long. Le domaine C2 peut se lier à une surface phospholipidique composée de phosphatidylcholine et de phosphatidylserine. Notre modèle du domaine C2 permet de proposer une orientation précise vis-à-vis de la membrane lipidique: 1) les chaînes latérales de résidus hydrophobes exposés au solvant sont enfouis dans la region hydrophobe de la membrane, 2) une couronne de résidus positivement chargés interagit avec la tête polaire des phospholipides et 3) l’axe principal du cylindre est perpendiculaire au plan de la membrane. Connaissant la structure des domaines A et C, notre objectif est de construire le modèle complet du factor Va.

III. Modèles moléculaires à haute résolution : assemblage de substrat dans son site recepteur

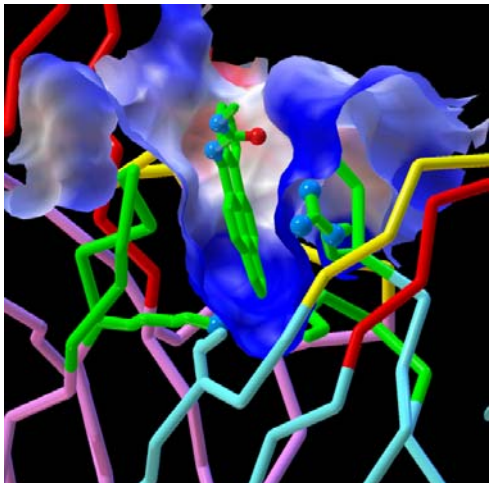
1. Modèle moléculaire de l'antigène spécifique de la prostate (Article no 3, annexe I).

Coombs G.S., Bergstrom R.C., Pellequer J.L., Baker S.I., Navre M., Smith, M.M., Tainer J.A., Madison E.L. & Corey D.R. (1998).
Substrate specificity of Prostate-specific antigen (PSA). Chem & Biol. **5**, 475-478.

L'antigène spécifique de la prostate (PSA), une protéase à serine, est un marqueur clinique très utile lors du cancer de la prostate. Le PSA est un membre de la famille des kallikreines, un sous-groupe de protéases à sérine, avec une différence néanmoins quant à sa spécificité qui est de type chymotrypsine au lieu du type commun trypsine. Notre travail consiste à définir la spécificité du PSA grâce à une stratégie basée sur le criblage de banques de phages couplée à la modélisation moléculaire du PSA et au docking de peptides substrats dans son site actif. Le criblage de banques de phages, incluant plusieurs cycles d'optimisation, révèle une séquence consensus peptidique de type SS(Y/F)Y|S(G/S) clivée avec une efficacité de l'ordre de 2200 à 3100 $M^{-1}s^{-1}$. Le positionnement du peptide substrat dans le site actif du PSA requiert un modèle moléculaire précis de cet enzyme. Un nouveau modèle du PSA a été construit à partir de la kallikreine de tissue et de la tonine. Ce modèle nous permet de positionner le peptide substrat dans une conformation dite consensus pour la famille d'inhibiteur/substrat de protéases à sérine. En particulier, le réseau de liaisons hydrogène entre les atomes du squelette peptidique du substrat (résidus P1 et P2) aux atomes du squelette protéique du PSA (résidus 193, 195, 214, 216) est strictement conservé. La tyrosine P1 (site de clivage) est capable de former une ou plusieurs liaisons hydrogène dans la poche de spécificité S1. La tyrosine P2 est partiellement enfouie sous la boucle 95-96 ce qui explique la forte tendance à trouver un résidu hydrophobe à cette position (Y ou L ou F). Récemment, des peptides ont été identifiés avec des clivages alternés (en position P1' au lieu de P1) provoqués par des mutations en positions P5 et P2. Nous allons utiliser notre modèle pour comprendre ce phénomène qui semble avoir une signification biologique dans la famille des kallikreines.

2. Modèle d'anticorps anti hydrocarbures aromatiques polycycliques.

Ce travail combine des analyses structurales à des étapes d'ingénierie de protéines dans le but de produire des anticorps capables de reconnaître, avec des affinités et des sélectivités variées, divers hydrocarbures aromatiques poly-cycliques (PAHs). A long terme, ce projet à pour but de développer une méthode d'immunodétection qui sera utile pour le dépistage et la surveillance quotidienne des PAHs *in situ*. Ces PAHs sont des molécules, de trois à six cycles benzènes, contaminantes de l'environnement, provenant principalement de la combustion de matériaux organiques, d'énergies fossiles ou encore par biosynthèse des plantes ou bactéries. Le principal défi pour la reconnaissance spécifique de PAHs provient de leur hydrophobie, leur planéité et leur manque d'atomes pouvant former des liaisons hydrogène. Deux fragments d'anticorps (Fab) recombinants qui réagissent de manière croisée avec le benzo-[a]-pyrène ainsi que d'autres PAHs ont été selectionnés. Nous avons construit un modèle tridimensionnel pour ces deux anticorps qui présentent un site de liaison à l'antigène particulièrement profond (figure 22) en raison d'une mutation du résidu très conservé Trp H47 par une valine ou une leucine. Cette poche, très polaire, est flanquée par deux chaînes latérales chargées (Lys et Arg) rendant possible des interactions de type π-cation entre l'anticorps et le PAH. Nous avons montré, grâce aux calculs de mécanique quantique et des interactions électrostatiques par la méthode FDPB, que la présence d'un groupe de liaison attaché au PAH (groupe nécessaire pour lier le PAH à une molécule porteuse) modifie fortement la distribution des charges partielles du PAH. Cet accroîssement de la polarisation du PAH justifie la présence de l'arginine enfouie dans la poche de reconnaissance de l'anticorps. Des expériences de mutagénèses dirigées sont en cours dans le but de comprendre l'induction de chaînes latérales chargées contre les PAHs.

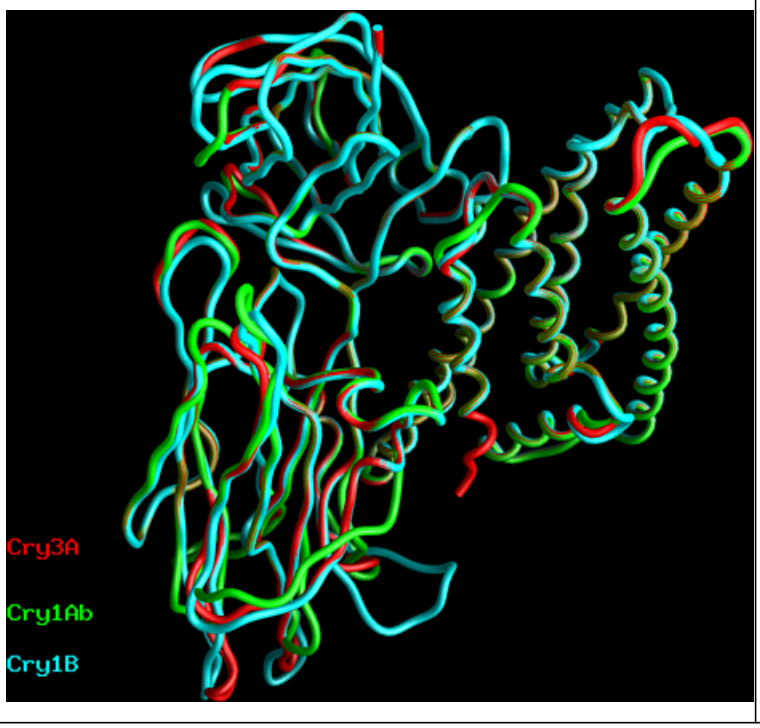


Site de reconnaissance d'un anticorps monoclonal anti-PAH. Le benzo-[a]-pyrène est positionné dans une poche profonde. La face des cycles benzène présentant un potentiel électrostatique négatif est en vis-à-vis de l'arginine du CDRH3. Le potentiel électrostatique de la poche de reconnaissance est fortement positif visualisé grâce au spectre coloré sur la surface moléculaire (bleu est positif, rouge est negatif, blanc est neutre). Figure construite par AVS.

Figure 22: Modèle du site de reconnaissance de l'anticorps anti-PAH 4D5.

IV. Ingénierie des protéines : développement d'immunotoxines.

Nous avons modélisé plusieurs toxines homologues de *Bacillus thuringiensis* (Bt). Cette bactérie produit des inclusions protéiques cristallines, lors de sa phase de sporulation, qui sont toxiques pour divers Lépidoptères, Diptères et des larves d'insects Coléoptères. Ces inclusions (δ-endotoxines) sont composées de plusieurs chaînes polypeptidiques qui, pour atteindre leurs formes actives, requièrent une activation par protéolyse dans les intestins d'insects. Deux toxines ont été modélisées à partir de la structure tridimensionnelle connue de Btt (Bt tenebrionis ou Cry3A) : Btk (Bt kurstaki ou CryIAb) et Cry1B (figure 23).



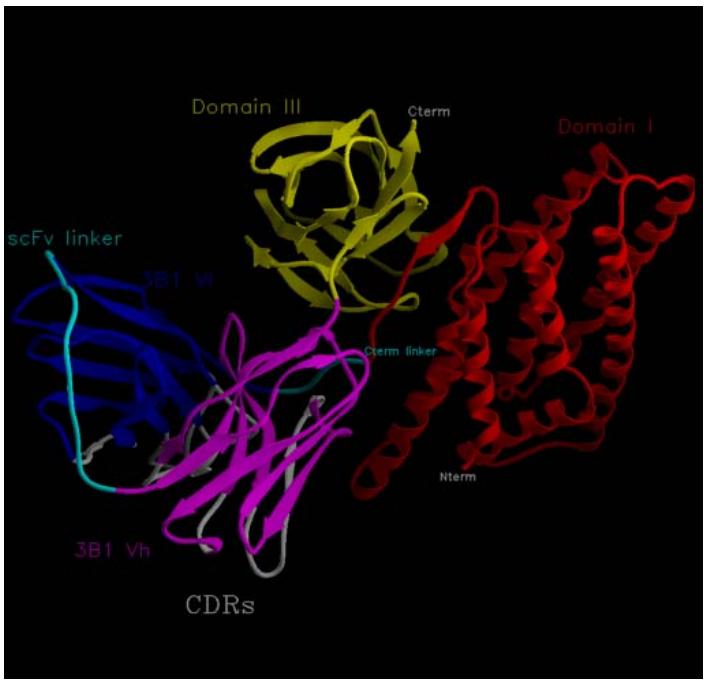
Modèles moléculaires des toxines de Bt : Cry1Ab (vert) et Cry1B (cyan) obtenues à partir de la structure connue de Cry3A (en rouge). Le repliement tertiaire est identique pour ces toxines à l'exception de boucles situées à l'extrémité du domaine II (bas de la photo), confortant l'hypothèse que le domaine II est impliqué dans la reconnaissance d'un recepteur. Figure construite par GRASP.

Figure 23: Modèles de toxines de *Bacillus thuringiensis*.

La structure tridimensionnelle de ces toxines est relativement conservée bien que leurs séquences divergent sensiblement. Ces toxines ont environ 600 acides aminés répartis en trois domaines consécutifs : le domaine I, entièrement en hélices α, est supposé s'insérer dans les membranes d'insects provoquant une fuite d'électrolytes et d'eau aboutissant à la mort en quelques minutes ; le domaine II, entièrement en feuillet β, a un repliement voisin d'anticorps à la seule différence que ce domaine est formé d'une seule chaîne. Ce domaine est supposé lier un recepteur membranaire non identifié à l'heure actuelle ; le domaine III, également complètement en feuillet β, forme un petit

globule dont le rôle n'est pas encore bien compris. La nature insecticidale de ces protéines présente beaucoup d'intérêt pour le développement de plantes transgéniques capables de résister aux attaques d'insectes tout en permettant d'alléger, voir de supprimer, l'usage d'insecticides chimiques.

La spécificité des toxines Bt est relativement étroite ce qui rend leur utilisation limitée. Ceci nous a conduit à développer des constructions d'immunotoxines (ou hybrides) où un ou plusieurs domaines de la toxine sont remplacés par des fragments variables (Fvs) d'anticorps que l'on a également modélisé. Le but est de pouvoir construire de manière rationnelle une toxine "sur mesure" avec une spécificité contrôlée fournie par le fragment d'anticorps. Nous avons réalisé 31 constructions différentes. Des efforts particuliers ont été employés pour produire l'immunotoxine suivante : Le domaine I et III sont conservés et le domaine II est remplacé par un Fv (figure 24). Cette construction semble logique puisqu'il est supposé que le domaine II est chargé de la reconnaissance spécifique du recepteur. Néanmoins, notre construction aurait été inutile si le domaine II n'avait pas une ressemblance proche avec un Fv autant du point de vue de la taille que de la forme. Le fait que le domaine II ait trois feuillets β alors qu'un Fv en a quatre, nous laisse supposer que ce domaine II est peut-être la forme ancestrale d'un Fv.



Construction d'une immunotoxine de Bt dans laquelle le domaine II a été remplacé par un fragment variable d'anticorps. Les conformations de quelques chaînes latérales, situées à l'interface des domaines II et III, ont du être modifiée pour obtenir ce modèle. On peut apprécier la similarité du domaine II et d'un Fv de part l'orientation similaire des boucles CDRs de l'anticorps comparées à celles du domaine II dans la figure précédente. Figure construite par RIBBONS et AVS.

Figure 24: Modèle d'une immunotoxine.

V. Modélisation moléculaire et bioinformatique : modélisation de protéines divergentes : la super-famille de sequences PAS (Article no 4, annexe I).

Pellequer J.L., Wager-Smith K.A., Kay S.A. & Getzoff E.D. (1998a).
Photoactive yellow protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily. Proc. Natl. Acad. Sci. USA **95**, 5884-5890.

La bioinformatique est la discipline qui tente de relier les séquences protéiques à leurs fonctions. Ici on présente une approche où grâce à une étude intensive des alignements de séquences, un prototype structural est défini pour une classe de domaines moléculaire appelés PAS (Per-Arnt- Sim, un acronyme pour les trois premières molécules identifiées dans cette famille). Les domaines PAS sont trouvés dans diverses protéines aux travers des trois règnes du vivant où leurs fonctions sont la détection et la transduction de signaux y compris la régulation de l'horloge biologique (protéines "Clock"). Bien qu' une grande quantité d'informations soient connues à la fois en séquence et en fonction, toutes ces données n'ont pas été intégrées au niveau structural. Le fait que les protéines contenant des domaines PAS ont évolué très tôt, tant au niveau de leurs séquences qu'au niveau de leurs fonctions, rendent leurs analyses complexes. Une homologie limitée de séquence, sur environ 50 résidus, a été identifiée entre les domaines PAS et la bactérie photosensible PYP (Photoactive Yellow Protein) dont on connaît la structure tridimensionnelle (3D). Cette homologie a été détectée grâce à l'utilisation de motifs de séquences consensus de la famille des phytochromes de plantes. Par contre la séquence de PYP ne peut détecter aucune homologie avec d'autres séquences PAS jusqu'à ce que trois résidus de PYP, spécifiquement choisis, aient été "mutés".

En calquant un domaine typique PAS, d'environ 150 residus, sur la structure 3D de PYP, on a pu montrer que les différences et similarités entre les domaines PAS et PYP sont en accord avec un repliement similaire. Un modèle moléculaire a été construit à partir d'une séquence PAS provenant du translocateur nucléaire du récepteur d'hydrocarbures aromatiques (ARNT), un acteur central dans la transduction de signaux intracellulaires. Le modèle confirme que PYP semble être le prototype structural pour les domaines PAS. De plus, le modèle est en accord avec les nouvelles observations qu'un ligand pourrait être lié au domaine PAS, pusiqu'une poche de taille raisonnable existe au centre du domaine PAS, reminiscente du chromophore présent dans PYP. Notre hypothèse de similarité structurale permet de proposer des expériences guides dans le but d'étudier l'éventuelle liaison d'un ligand, la dimérisation du domaine PAS et la transduction de signaux cellulaires. Récemment notre modèle moléculaire a été confirmé par une structure cristallographique (Cabral et al. 1998) ainsi que notre hypothèse de ligand (Gong et al., 1998).

CONCLUSIONS

La modélisation moléculaire devient de plus en plus une discipline intégrée aux domaines de la biologie expérimentale. Son rôle est de fournir aux biologistes les images “atomiques” de la conformation tridimensionnelle des objets qu’ils étudient. Nous avons présenté en détail la technique de modélisation par homologie. Cette technique permet de prédire la conformation d’une protéine grâce à son homologie de séquence avec des protéines ayant leurs structures tridimensionnelles déterminées expérimentalement.
La modélisation par homologie comprend plusieurs étapes (reconnaissance de repliement, alignement de séquences, construction de la chaîne principale et des chaînes latérales, construction des insertions et suppressions, affinement). Il a été établi que l’alignement de séquences et la construction des insertions/suppressions sont les étapes limitantes. Ces deux étapes dépendent l’une de l’autre. Un mauvais alignement de séquence provoquera une mauvaise modélisation des insertions/suppressions. La raison principale de la limite rencontrée dans l’alignement de séquences est due au fait que, localement, l’homologie de séquences n’est pas significative. Ironiquement, cette limitation ne peut être résolue uniquement au niveau tridimensionnel. Peut-être est-il possible d’affiner ou de tester un modèle moléculaire en modifiant l’alignement de séquences systématiquement ? En ce qui concerne la construction des insertions/suppressions, il semble que les efforts doivent se concentrer sur les techniques d’échantillonnage.

La modélisation par homologie est une technique de prédiction qui a un large spectre d’applications. Nous avons présenté quatre exemples d’applications dans ce mémoire : 1) Etude des mécanismes fonctionnels, 2) Assemblage de substrat/ligand dans leur récepteur, 3) Ingénierie des protéines et 4) Bioinformatique.
La première application concerne la visualisation du repliement des protéines impliquées dans un processus biologique. L’exemple présenté fut le modèle du cofacteur Va de la cascade de coagulation. Ce modèle a permis non seulement de proposer un mode d’ancrage du cofacteur Va à la membrane phospholipidique, mais aussi de montrer le mécanisme de régulation de ce cofacteur par son assemblage avec la protéine C activée. Cette application est le prototype même de l’utilisation de la modélisation moléculaire dans le but de proposer une lignée d’expériences à réaliser.
La seconde application, très convoitée par les industries pharmaceutiques, concerne la prédiction qualitative et quantitative de l’assemblage d’un ligand dans son site récepteur. Elle requiert des modèles moléculaires à hautes résolutions qui ne peuvent être obtenus uniquement lorsque les homologies de séquences sont importantes. A l’heure actuelle, seules les familles de protéines bien étudiées expérimentalement et struc-

turalement peuvent produire de tels modèles (Anticorps, protéases à sérine, repliement à quatre hélices...). Nous avons présenté la modélisation de l'assemblage du substrat d'une protéase à sérine, l'antigène spécifique de la prostate. La précision de notre modèle permet d'expliquer quantitativement la préférence séquentielle du substrat, à la position P1, grâce à un simple calcul d'énergie potentielle, en montrant qu'un résidu aromatique et polaire était préférable à un résidu court et apolaire. Nous avons également présenté l'assemblage d'un haptène (benzo-a-pyrene) dans un site de reconnaissance d'un fragment variable d'anticorps. Le but de cet assemblage est de pouvoir prédire quelles modifications l'anticorps doit subir pour pouvoir reconnaître spécifiquement un unique hydrocarbure aromatique polycyclique, caractérisé par le nombre de cycles benzènes. Ces études sont caractéristiques de la modélisation moléculaire où les techniques de modélisations par homologie, d'assemblage intermoléculaires, de calculs de mécanique quantique sont employées synergiquement pour atteindre le but désiré.
La troisième application est certainement la plus ambitieuse : l'ingénierie de protéines. La problématique fut la suivante : on dispose de plusieurs toxines de Bacillus spécifiques de certains insectes nuisant aux récoltes agricoles ; cependant aucune de ces toxines ne protègent la plante étudiée. Notre objectif fut de proposer le remplacement du domaine de la toxine responsable de la liaison à un récepteur spécifique par un fragment variable d'anticorps spécifique d'un ou plusieurs récepteurs présents à la surface des intestins d'insectes cibles. La difficulté de la modélisation dans ce cas est d'être certains que le modèle généré ne s'oppose ni au repliement ni à la stabilité de la toxine.
La quatrième application créée un lien entre la modélisation moléculaire et la bioinformatique. Ce lien est l'identification de la fonction d'un domaine d'une protéine. L'exemple que nous avons présenté est celui des domaines PAS. Aucune fonction spécifique ne leur est connue et l'homologie de séquences est bien au dessous du seuil de significativité. La modélisation par homologie a permis de montrer que le repliement de la protéine PYP peut accomoder les séquences de plusieurs domaines PAS. Par analogie avec PYP il est postulé que les domaines PAS sont aussi impliqués dans des processus de signalisations. Des résultats récents semblent confirmer ces hypothèses puisqu'une structure cristalline montre qu'un domaine PAS peut lier un groupement hème (protéine FixL).

L'avenir de la modélisation par homologie est prometteur. Les points faibles ont été identifiés et ne dépendent uniquement du pourcentage d'homologie de séquences entre deux protéines. L'apport de nouvelles structures tridimensionnelles expérimentales devrait réduire l'espace entre les molécules cibles et les molécules à structures connues conduisant inexorablement à une amélioration des prédictions.
Par contre l'avenir de la modélisation moléculaire est incertain. Il est fortement couplé à son succès à prédire la conformation de complexes intermoléculaires. Cette tâche est compromise en raison de la flexibilité intrinsèque des protéines ainsi que leur flexibilité

extrinsèque provoquée par l'interaction avec une autre molécule. En effet, le plus grand défi de la modélisation moléculaire est de pouvoir prédire et contrôler un changement conformationnel. Un axe de recherche priviligié sera celui de l'analyse des structures tridimensionnelles dans le but d'en extraire leur "substantifique moelle". En effet, malgré presque bientôt 9000 structures de protéines (RX et RMN) dans la base de données PDB, notre compréhension du changement conformationnel n'en est qu'à son balbutiement. Il sera nécessaire d'employer des techniques expérimentales d'analyse de changements conformationnels comme la RMN ainsi que des techniques fonctionnelles comme la mesure de constantes d'affinité.

Notre orientation scientifique à long terme est d'utiliser la synergie des approches théoriques et expérimentales dans le but d'étudier les interactions responsables de la reconnaissance intermoléculaire. L'important est de travailler sur un système modèle qui se prête parfaitement aux techniques d'analyses expérimentales ainsi qu'aux techniques informatiques. Le meilleur système d'étude de la reconnaissance intermoléculaire est celui des antigènes-anticorps, mais d'autres systèmes sont envisageables comme par exemple le second domaine des toxines de Bacillus. Dans l'immédiat, nous allons porter nos efforts sur deux systèmes : les anticorps anti-PAH et les anticorps anti-PCB. Nous disposons des systèmes d'expression pour ces deux anticorps ainsi que des tests de mesures d'affinité. Notre première étape sera de cristalliser ces deux anticorps et de résoudre leurs structures cristallographiques en complexes avec leurs ligands respectifs. Les familles des hydrocarbures aromatiques polycycliques et des PCBs sont suffisamment larges pour étudier les changements conformationnels éventuels lors de réactions croisées. La combinaison en parallèle des techniques de modélisations par homologie aux techniques cristallographiques porte beaucoup d'espoirs dans notre quête d'analyses des changements conformationnels. Le retour au monde expérimental après cinq années d'études informatiques est certes une grande satisfaction !

ANNEXE I : Publications

ANNEXE II

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Bovine α-lactalbumin	Lysozyme			INT	MEC				Browne et al., 1969
Trypsin, elastase	α-chymotrypsin	46,42		INT	MEC				Hartley, 1970
Myxobacter 495 α-lytic protease	Elastase, α-chymotrypsin			INT	MEC				McLachlan & Shotton, 1971
Troponin C	EF hand of parvalbumin		157	INT	MEC				Kretsinger & Barry, 1975
Relaxin	Porcine insulin		~51	INT	MEC				Bedarkar et al., 1977
Relaxin	Insulin		~51	INT	MEC				Isaacs et al., 1978
Human insulin-like growth factor	Porcine insulin		51	INT	MEC	MF			Blundell et al., 1978
Haptoglobin heavy chain (HpH)	Serine proteases (x3)			INT (3 models)		PA			Greer, 1980
Bovine FXa	Elastase			INT				Prothrombine clivage peptide	Greer, 1981b
β-crystallin (βBp)	γ-crystallin (γII)		178	INT	FR			Dimer of βBp	Wistow et al., 1981
Casiragua insulin	Porcine insulin		~51	INT	FR				Blundell & Horuk, 1981
Shark, procine relaxin	Insulin			INT	FR				Bedarkar et al., 1982
Factor Xa, IXa, thrombin	Chymotrypsin, trypsin	~50	~245	INT/ SCR-mix	Graphical display PS				Furie et al., 1982
Rat relaxin	Insulin 4Zn		~51	INT					Dodson et al., 1982

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Murine β-crystallin b23	Bovine γ-II crystallin		183	INT	FR/ MI				Inana et al., 1983
Urokinase (UK), tissue-type plasminogen activator (TPA)	Chymotrypsin		~250	INT/ SCR	FR				Straßburger et al., 1983
Mouse submaxillary renin	Endothiapepsin			INT	FR				Blundell et al., 1983
Human renin	Endothiapepsin		327	INT	FR				Sibanda et al., 1984
Human HLA-D region	Constant region of Fab NEW		~210	INT	FR				Travers et al., 1984
E. coli glutaredoxin	*C. nephridii* / *E. coli* / *T4* thioredoxins	47/ 47/24	~91	INT	FR				Eklund et al., 1984
Human renin	Penicillopepsin	25		INT	MEC				Akahane et al., 1985
Human renin	Rhizopuspepsin	30	339	INT	FR	CH			Carlson et al., 1985
Calmodulin	Carp parvalbumin, intestinal calcium-binding protein	H≈20		INT	CHE	AM			O'Neil & DeGrado, 1985
Human renin	Aspartic proteinases (x4)			INT/ SCR		VF			Plattner et al., 1986
Rhodopsin	Bacteriorhodopsin			INT					Findlay & Pappin, 1986
Angiogenin	Ribonuclease	H≈35	118	INT	AVP	EC			Palmer et al., 1986
Human activated protein C inhibitor	α1-antitrypsin	43	~390	INT	FR	EC		APC	Toma et al., 1987
Bombyx mori prothoracicotropic hormone (PTTH-II)	Insulin		~51	INT	FR				Khoti et al., 1987
HIV retroviral pol-protease	Endothiapepsin		~300	INT	FR				Pearl & Taylor, 1987

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
α-Lactalbumin	Hen egg white lysozyme		123	INT		LU			Robson & Platt, 1987
Frog lens βA1-crystallin	Bovine γ-UU crystallin	38	198	INT	FR/ MI				Luchin et al., 1987
Bacteriophage IKe DNA binding protein (IKe-DBP)	Bacteriophage fd BP(G5BP)	44	80	INT			0.43 core		Brayer, 1987
Chloramphenicol acetyltransferase (CAT)	Cat muscle pyruvate kinases			INT		LU			Robson et al., 1987
Human growth hormone	de novo			Biophysical data					Cohen et al., 1987
α-subunit of tryptophane synthase	de novo			Biophysical data					Hurle et al., 1987
Rhodobacter sphaeroides Y thioredoxin	*E. coli* thioredoxin	49	107	INT	FR				Clement-Metral et al., 1988
Human renin	Pepsin, penicillopepsin	21-70		INT/ SCR		VF		Angiotensinogen analogues	Sham et al., 1988
Human CD4 Nterm binding domain	VL REI		~178	INT/ Spare	QU	CH			Bates et al., 1989
HIV-1 protease	Rous sarcoma virus protease	~24	99	INT					Weber et al., 1989
Human Nterm intercellular adhesion molecule 1 (ICAM-1)	Ig constant domain		84	INT	FR			Human rhinovirus-14 (HRV14)	Giranda et al., 1990
Amphioxus calcium vector protein (CAVP)	Calmodulin, troponin C	49, 47	143	INT	BRU				Cox et al., 1990
Human P450 17a	P450cam	13	467	INT/ Spare parts	QU	AM			Laughton & Neidle, 1990
B-type chymopapain	Papain			COM		GR			Topham et al., 1990b

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Papaya proteinase omega	Papain	72		COM	SY	SY			Topham et al., 1990a
Bacillus neutral protease (NP)	Thermolysin	H=50	300	INT	FR				Signor et al., 1990
G-domain chloroplast elongation factor Tu ($EF\text{-}Tu_{chl}$)	*E. coli* EF-Tu	H=70	175	INT	MMS	AM		GDP	Lapadat et al., 1990
RNase (Pch1, Ms)	Aspergillus RNase T1	70-60	~104	INT	IN	DI	**1.89 bk**		Floegel et al., 1990
Human A1 hnRNP RNA binding domain	Horse muscle acylphosphatase (HMA)	H=25 %	93	INT	FR				Ghetti et al., 1990
Donor side components in photosystem II (PSII)	Phostosynthetic reaction center			INT	FR				Svensson et al., 1990
Human C5a anaphylatoxin (7 helix bundles)	Bacteriorhodopsin	5-60	~170	WI		DI			Grötzinger et al., 1991
Hamster aspartate transcarbamylase (ATC)	*E.Coli* ATC	44	~310	INT	QU	CH	0.93 Cα		Scully & Evans, 1991
Human monocyte chemoattractant and activating protein (MCAF/ MCP-1)	Interleukin-8	24	76	INT	QU	CH/ XP	0.84 bk		Gronenborn & Clore, 1991
Human Myb (DNA binding domain)	de novo		~50	DM/ INT	IN/ FR	DM			Frampton et al., 1991
Human cytochrome P450 (IA1)	*Pseudomonas putila* P450cam	15	~469	INT/ Spare parts		CH		Heme	Zvelebil et al., 1991
ATP-binding domain of permeases	Adenylate kinase		~260	INT	IN				Mimura et al., 1991
Papaya proteinase omega	Papain, actinidin	72.4, 52.3	216	COM/ SCR	FR	GR			Topham et al., 1991
Methanococcus ferredoxin (FdMt)	Peptococcus ferredoxin (FdPa)		60	INT	TOM-FR	XP			Bruschi et al., 1991

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Homarus gammarus crustacyanin C1 (CRTC)	Retinol-binding protein	38	181	INT					Keen et al., 1991
Peptide binding domain of hsp70	Binding domain of HLA		~180	INT	QU				Rippmann et al., 1991
Reverse transcriptase ribonuclease H (from MuLV, HIV)	*E. coli* RNase H	H<30	~150	INT/ FRG	FR	PR/ AM	1.4 bk		Nakamura et al., 1991
Rhus vernicifera stellacyanin	Cucumber basic protein	39	107	INT		XP	~1.8 bk		Fields et al., 1991
Human and rat amylin	α-calcitonin gene-related peptide (α-CGRP)	46	37	INT	QU	CH			Saldanha & Mahadevan, 1991
Human pancreatic secretory trypsin inhibitor	Porcine PSTI	73		DIS		AM	1.0 Cα		Havel & Snow, 1991
Human interleukin-4	de novo			Biophysical data					Curtis et al., 1991
Phlebia radiate lignin peroxidaseLIII	Cytochrome C peroxidase	21.1	~337	COM		CH		heme	Hoffrén et al., 1991 ; Hoffrén et al., 1993
Rat liver formaldehyde dehydrogenase (FALDH)	Horse liver alcohol dehydrogenase (EE-ADH)	63	374	COM/ INT	FR				Lapatto, 1991
Human thyroxine binding globulin (TBG)	α1-antitrypsin	H=42	395	INT	IN	DI		Thyroxine	Jarvis et al., 1992
Human TBG	α1-antitrypsin	H=42	395	INT	FR			Thyroxine	Terry & Blake, 1992
Rana pipiens P-30 protein (onconase)	Bovine pancreatic RNase A	28	104	MUT/ INT	TOM	CH			Mosimann et al., 1992

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Schizosaccharomyces pombe pyrophosphatase PPA2	PPA1	~50	~323	INT	SY	AM			Vihinen et al., 1992
Bovine cyclic GMP-binding domain of cyclic GMP-gated ion channel (cGGC)	Cyclic AMP binding domain of catabolite gene activator protein (CAP)	17	118	INT	FR	XP	0.53 bk		Kumar & Weber, 1992
E. coli tyrosine aminotransferase	*E. coli* aspartate aminotransferase	42.9	405	INT		XP/ GR	1.29 Cα		Jäger et al., 1992
Carcinoembryonic antigen (CEA)	Ig, CD2, CD4 domains			INT	QU	CH			Bates et al., 1992
Fifth domain of human β2-glycoprotein I	16th repeat of factor H	15	~65	INT	IN	XP			Steinkasserer et al., 1992
Saccharomyces cerevisiae elongation factor 2	EF-Tu	H=55	161	INT	UCS D soft	XP	0.79 bk		Perentesis et al., 1992
Pisum sativum photosystem II reaction center	Rhodopseudomonas and rhodobacter PSII		~344	COM		SY			Ruffle et al., 1992
Basic cytosolic bovine 21 kDa protein	Yeast phosphoglycerate kinase	20.4	186	INT	MA/ IN	CH			Schoentgen et al., 1992
Lignin peroxidase	Cytochrome C peroxidase	21.4	343	INT	MI/ QU	AM			Du et al., 1992
G-protein-coupled receptors	Bacteriorhodopsin			INT	SY			Neurotransmitters (INT/MD)	Trumpp-Kallmeyer et al., 1992
Bacillus alcalophilus alkaline subtilisin	Subtilisin Carlsberg	H≈60	269	BRA/ INT	BRA	AM	**1.21 bk**		Aehle et al., 1993 ; Aehle et al., 1995
Pyrococcus furiosus rubredoxin	Clostridium (x1), Desulfovibrio (x2)	H≈49 -57	~65	INT	SY	AM	1.15-1.96 bk		Wampler et al., 1993

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Mouse submandibular gland serine protease subunits (γ-NGF, α-NGF, EGF-BP)	Glandular kallikrein, tonin	53-64	~245	COM/ INT	SY/ FR	CO (CC P4)		EGF	Bax et al., 1993
Mouse mast cell chymases (mMCP-1,2,4,5)	Serine proteases (x9)	55-75	226	MOD /SCR		CH	0.2 Cα		Sali et al., 1993
Calcineurin B	Calmodulin	35	~169	INT	SY	CH			West et al., 1993
E-selectin endothelial-leucocyte adhesion molecule (ELAM-1)	Rat mannose-binding protein	27		COM/ INT/ SCR	FR	SY-TR			Mills, 1993
Human and murine low affinity receptor for IgE (fcεRII/ CD23)	Rat mannose-binding protein	30	113	INT	FR	XP			Padlan & Helm, 1993
Human β2-adrenoreceptor (β2adr)	Bacteriorhodopsin	~9		INT/ Biophysical data	IN/ WI	DI			Cronet et al., 1993
Human interferon α-2 (HuIFNα2)	Murine interferon β		165	STE/ INT	SY	IM			Murgolo et al., 1993
Gonadotropin releasing hormone	de novo		10	DB search	QU	CH			Gupta et al., 1993
Human cathepsin D	Pepsin, chymosin	50.6, 47.8		COM/ SCR-mix	FR	SY	0.52-1 Cα	Synthetic peptide	Scarborough et al., 1993
Ion channel of Influenza virus M2 protein	de novo			BND LQ	QU	CH			Samsom & Kerr, 1993
Coiled-coil α-tropomyosin	Cα atoms			INT	FR/ IN	AM			Cregut et al., 1993
Muscarinic m1 receptor	Bacteriorhodopsin			INT	SY	AM		Muscarinic agonist (INT/ min)	Nordvall & Hacksell, 1993

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
α1 antichymotrypsin (ACT)	Ovalbumin	29	377	INT	IN	XP	2.4 Cα	P3-P3' chymotrypsin peptide (TRA)	Katz & Christianson, 1993
Human IL4 receptor, IL4	Human CD4, granulocyte macrophage colony stimulating factor (GM-CSF)	19.4,?	191, ~129	COM/ INT	SY	CH	**2.43, 1.86** Cα **(core)**	IL4-IL4 receptor (INT/TRA/ min/MD)	Bamborough et al., 1993
Biotinylated domain of yeast pyruvate carboxylase, lipoylated H-protein of pea leaf glycine cleavage system	Bacillus lipoyl domains, *E. coli* PDH		~69, ~74	DC	QU	XP			Brocklehurst & Perham, 1993
α-conotoxin GI, endothelin	de novo			Random search on ϕ,ψ.		EC	**~4 bk**		Sowdhamini et al., 1993
E. coli flavodoxin	Flavodoxins	16-44		DC/ SCR	IN	DI	1.52 all		Havel, 1993
Murine Il-4	Human IL-4			DC			0.64 bk		Srinivasan et al., 1993
Human complement enzyme factor D	Rat mast cell protease	33-36	228	INT	ATOM	XP	1.65 xray		Carson et al., 1994
Activated protein C (APC)	Chymotrypsin	32	~419	INT/ SCR	IN	DI	1.07 bk (SCR)		Fisher et al., 1994
Human furin	Subtilisin BPN', thermitase	24,29	330	INT/ SCR-mix	QU	CH		Substrate peptide	Siezen et al., 1994
Streptococcus mutans ingbritt (HPr)	*Bacillus subtilis* HPr (x2)	62.1	87	HOM /INT/ SCR	IN	AM	0.21 bk		Dashper et al., 1994
Human proline-rich homeodomain protein Prh	MAT α2 homeodomain	H=28	61	INT/ SCR	MI	AM			Neidle & Goodwin, 1994

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Rat cyt b5 heme-binding domain	Bovine cyt c5	93	~98	INT	QU	CH	1.28 Cα (core)		Gill et al., 1994
Abl-SH3 domain	Human Fyn-SH3	47	57	INT	IN	DI	0.58 Cα		Pisabarro et al., 1994
C4b-binding protein (C4BP) b chain	Factor H		60	HOM /INT	IN	DI			Fernán-dez et al., 1994
Lysosomal protective protein/ Carboxypepti-dase L (CARB L)	Wheat carboxypepti-dase II	32	~400	INT	BI	DRII	1.01 Cα		Elsliger & Potier, 1994
α-hordothionin	Crambin	29	46	HOM	IN	DI	0.78 bk		Rao et al., 1994
Pneumocystis carinii dihydro-folate reductase binding site	*L. casei* DHFR	H=80	206	INT	QU	CH	0.64 bk	Methotrex-ate (MTX, TRA)	Souther-land, 1994
Dopamine D2 receptor	Bacterior-hodopsin	~9		INT	FR			Apomor-phine, cyprohep-tadine (INT)	Teeter et al., 1994
Human throm-boxan A2 (TXAS)	P450bm3	26.4		INT	QU			Prostag-landin H2 (TRA/min)	Ruan et al., 1994
Annexin I	Annexin V	43		Corea 90/ HOM / SMD/ LECS	IN	AM	**1.81-2.04 bk**		Cregut et al., 1994
Human class Mu (M1b-1b, M2-2, M3-3) glu-tathione S-transferases	Rat3-3 GST isozyme	60-79		INT/ Den-sity restr aints	IN	XP	0.6-1.15 bk		Cachau et al., 1994
Human PSA,	Kallikrein	55		HOM /INT/ SCR	IN	DI	0.7 bk	P3-P3' peptide substrate (TRA)	Villoutreix et al., 1994
Human PSA, human glandu-lar kallekrein (hGK)	Porcine pan-creatic kal-likrein A	59.6, 64.3		INT	SY				Vihinen, 1994

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Class II histocompatibility protein HLA-DR1	Class I MHC	~12	179	QU/ design	QU	CH	**1.49, 4.33** Cα		Nauss et al., 1995
Type II antifreeze protein (AFP)	Rat mannose binding protein (MBP-A)	19	129	HOM /INT	IN	AM	0.8 bk		Sönnichsen et al., 1995
Human complement protease C1s	Serine protease, complement factor H		~409	INT/ SCR/ Cross-linking exp	O				Rossi et al., 1995
Human O^6-alkylguanine-DNA alkyltransferase (hAT)	*E. coli* Cterm domain Ada protein	27.5	207	INT	QU	CH	0.9 Cα		Wibley et al., 1995
Rhodospirillum molischianum light-harvesting complex II (LHII)	Multiple proteins		~101	MOD /INT	QU	XP	**2.9** Cα		Hu et al., 1995
Saccharomyces cerevisiae domain III α-agglutinin	Ig KOL	13	~120	HOM /SCR/ INT	IN				Lipke et al., 1995
Drosophila domain of fasciclin III	VL McPC603	20/ SCR	103	HOM /INT	IN	DI			Castonguay et al., 1995
Phanerochaete chrysosporium Mn peroxidase	Lignine peroxidase	86	358	INT		DI			Selvaggini et al., 1995
Dianthin 30 (Type 1 ribosome inactivating protein-RIP)	Pokeweed antiviral chain A of type 2 RIP		~271	INT/ SCRave	QU	CH			Bravi et al., 1995
Thymocyte Thy-1	VL RHE			HOM /INT	IN	AM			Renouf & Hounsell, 1995
Lactobacillus bulgaricus D-lactate dehydrogenase	Pseudomonas formate dehydrogenase	22	332	HOM /INT	IN	DI			Vinals et al., 1995

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
GroES	de novo			LINUS					Srinivasan & Rose, 1995
Phospholamban transmembrane domain	de novo			Exp. data/ Rigid body search		XP			Adams et al., 1995
Human estrogen receptor	Human α antitrypsin	11	595	INT	SY	TR		Estradiiol	Lewis et al., 1995
Thiobacillus ferrooxidans rusticyanin	Nitrite reductase, blue copper proteins	19.4	155	INT/ SCR	IN	XP			Grossmann et al., 1995
Mouse mast cell protease 7 (mMCP-7)	Bovine pancreatic trypsin	39	245	MOD /INT		CH	~0		Matsumoto et al., 1995
Thermotoga maritima D-glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	3GAPDHs	>50	173	INT/ SCR-mix	IN	DI			Szilágyi & Závodszky, 1995
Lactococcus lactis leader peptidase NisP	Subtilisin BPN', thermitase	30	283	INT/ SCR-mix	QU	CH		P8-p2' substrate nisin peptide (TRA/ min)	Siezen et al., 1995
Cytochrome P450 2B1	P450s		~491	CONS/ INT	IN			Androsterredin, progesterone (INT/ CONS search)	Szklarz et al., 1995
Sox-5 high mobility group (HMG)-box protein	HMG1 B-box (NMRs)	29		Cluster of NMR models	MI	AM			Adzhubei et al., 1995
PSA, hGK (hK2)	Porcine kallikrein + rat tonin	H=53 .6, 67		INT/ SCR			1.3, 1.4 Cα		Bridon & Dowell, 1995
A domains of FVIII	Ceruloplasmin, nitrite reductase	11-28	987	HOM	IN	DI			Pan et al., 1995

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Halorhodopsin, Rhodopsin	Bacteriorhodopsin	34	237,282	INT/ Scat. factors	SY	GR	3.3, 5.1 Cα		Neumüller & Jähnig, 1996
Domains D1/D2 of Synechocystic photosystem II (PSII)	Bacterial reaction center	20-24	~698	INT/ SCR	QU	CH	0,82 bk		Xiong et al., 1996
Rat Mu class glutathione S-transferase 4-4	Rat GST3-3	78	434	INT	QU	CH	0.85 bk	GST conjugate (INT)	de Groot et al., 1996
Rabbit liver microsomal P450	Cytochromes P450	18		INT/ SCR	QU				Chang et al., 1996
Nterm zinc ring domain of breast anti-ovarian cancer succeptibility gene (BRCA1)	Equine Herpes virus zinc ring domain	38	79	HOM	IN	DI	3.0 bk		Bienstock et al., 1996
Human rap-1A protein	ras-gene encoded p21 protein	H=80	168	INT	SY		0.7 bk		Chen et al., 1996a
Rat upstream binding factor (rUBF)	Rat HMG box			HOM /EM analysis	IN	DI			Neil et al., 1996
Streptomyces cytochrome P450 ChaP (CYP105C1)	P450eryF	37	~381	HOM /INT/ SCR	IN/ QU	AM			Chang & Loew, 1996
Bacillus circulans xylanase	Superoxide dismutase	Thre ading	185	HOM /INT	IN	DI	**14.8** Cα		Chen et al., 1996b
E. coli acetohydroxyacid synthase (AhAS)	Lactobacillus pyruvate oxidase	27	548	INT	FR/O	DI		Thiamin pyrophosphate (TPP)	Ibdah et al., 1996
Human androgen receptor	Rat glucocorticoid receptor			INT	IN	AM		Oligonucleotide	Lobaccaro et al., 1996
Rat luteinizin hormone/chorionic gonadotropin receptor (LH/CG-R)	Porcine ribonuclease inhibitor (LRRs)		217	INT	FR/O	SCU LPT			Bhowmick et al., 1996

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Human methylmalonyl-CoA mutase (MCM)	*P. shermanii* MCM	65		MOD	QU				Thomä & Leadlay, 1996
Human annexin I	Annexin V	41	74	INT		CHARM	**0.32, 1.10 bk**		Musat et al., 1997
Herpes simplex virus I thymidine kinase (HSV1 TK)	Porcine adenylate kinase (ADK)			COM/INT		AM	**70%**	AMP/ATP	Folkers et al., 1997
Human complement protease C1r	Serine proteases		~409	INT/SCR/Cross-linking exp	O				Lacroix et al., 1997
Measles virus receptor Cd46	Factor H			INT/DIAM		EC/2			Mumenthaler et al., 1997
Rat submaxillary Kallikrein	Rat tonin, porcine kallikrein	74	295	INT	QU	CH			Henriques et al., 1997
Cystic fibrosis transmembrane conductance regulator (NFB1 domain)	ATPase F1			DIA/INT		XP			Annereau et al., 1997
Transmembrane helices of opioid receptor	Rhodopsin			de novo					Strahs & Weinstein, 1997
Antimicrobial protein Ace-AMP1	Non-specific lipid transfert proteins (ns-LTP)	26-33	93	MOD		XP	3.1-3.3 Cα		Gomar et al., 1997
Murine homeodomain MSX-1	Engrailed homeodomain protein	48	60	CONG		CH			Li et al., 1997
Saccharomyces cerevisiae adenylsuccinate synthetase	*E. coli* AS	~30	433	INT	SY	XP	0.47 bk, 1.42 after MD		Sticht et al., 1997
Human P450 2D6 (CYP2D6)	P450bm3 (CYP102)	20.8	453	INT	SY	TR		Ligands and substrates	Lewis et al., 1997

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Human P450 3A4 (CYP3A4)	P450bm3 (CYP102)	27.1	503	INT	SY	TR		Substrates	Lewis et al., 1997
Neurospora crassa geranylgeranylpyrophosphate synthase (GGPPS)	Avian farnesyl PPS	21	433	HOM /INT	IN	DI			Quondam et al., 1997
Rabbit cytochrome P450 2B4	P450s	<20	~491	INT/ SCR-mix	IN	AM		Benzphetamine, androstenedione (INT)	Filizola et al., 1997
Human cytochrome P450 1A2	P450bm3	21		HOM /WI/ INT	IN	WI/ DI		Caffeine, carcinogenic aromatic amine (MeIQ, autodock)	Lozano et al., 1997
Pyrococcus furiosus pyrolysin, *Thermococcus stetteri* stetterlysin	Subtilases	~44	~509, ~488	INT	QU	CH		Precurosr pyrolysin peptide (P4-P2', TRA)	Voorhorst et al., 1997
Cytochrome P450 3A4	P450s	~20	~491	CON S	IN			Progesterone, erythromycin (INT/MD)	Szklarz & Halpert, 1997
Ligand-gated ion-channel (LGIC) glycine receptor α1 subunit	SH2 (domain A), SH3 (domain B)	23 (dom A)	137 (A), 63 (B)	HOM	IN	AM		Strychnine (INT)	Gready et al., 1997
A domains of FVIII	Ceruloplasmin	34		HOM	IN				Pemberton et al., 1997
Human Cdc25 phosphatase catalytic domain	Rhodanese	17	122	ICM	TF		**0.2 core , 1.6** Cα		Hofmann et al., 1998
Leucine-rich repeat	Ribonuclease inhibitor		~30	INT	TF	XP			Kajava, 1998
P450: CYP11, CYP17, CYP19, CYP21	CYP102							Substrate inhibitor	Lewis & Lee-Robichaud, 1998

Cible	Parent	% Id	Taille	Meth	Graf	Min	RMS	Docking	Auteurs
Entamoeba histolytica ferredoxin (EHFXD)	Clostridium and azotobacter FXD	H≈high	59	HOM/INT	IN	DI			Mukhopadhyay & Lohia, 1998
Rhodobacter spheroides PYP	*E. halophila* PYP	48	124	WI/INT		GR			Kort et al., 1998
Viscum album mistletoe lectin I	Castor bean ricin	41 (A), 64 (B)	~530	INT	FR/WI				Eschenburg et al., 1998
C4b binding protein	Factor H		491	Fernández94	IN	DI			Villoutreix et al., 1998
Hamster galectin-3	Bovine galectin-1		135	LOO	O		0.59 Cα (core)	Galactose (TRA)	Henrick et al., 1998
Barley low-temperature-inducible gene family (BLT4)	Maize lipid transfert protein (LTP)	65-68	90	MOD	QU			Palmitic acid, octadecandienoyl-PG/PC (INT/MD)	Kereszt-essy & Hughes, 1998
PAS domain of ARNT	PYP			INT	XF/TF/IN	XP			Pellequer et al., 1998a
Human prostate specific antigen	Kallikrein	60	237	INT	XF/TF/IN	XP	0.075	Substrate peptide P4-P2'	Coombs et al., 1998
A domains of coagulation factor Va	Cerruloplasmin	39	1022	INT	XF/TF/IN	XP			Pellequer et al., 1998b
C domains of coagulation factor Va	Binding domain of galactose oxidase	13	319	INT	XF/TF/IN	XP	1.0	Phospholipids	Pellequer et al., 1998c
Coagulation factor Va	A and C domains of FVa		1341	INT	XF/TF/IN	XP		APC, phospholipids	Pellequer99

Première colonne (CIBLE) : nom de la protéine modélisée (cible).
Deuxième colonne (PARENT) : nom de la (ou des) protéine(s) parente(s).
Troisième colonne (% Id) : identité de séquence, si connue. Lorsque seulement l'homologie est connue,

le nombre est précédé par la lettre H.
Quatrième colonne (TAILLE) : taille de la protéine modélisée. Le signe ~ signifie que la taille n'est pas précisée explicitement et provient de l'alignement de séquences présent dans l'article.
Cinquième colonne (METH) : méthode de modélisation : COM (COMPOSER), CONG (CONGEN), CONS (consensus de plusieurs structures parentes), de novo (indique l'utilisation de techniques biophysiques couplées aux connaissances structurales), DC (contraintes de distances), DIA (DIANA), DIAM (DIAMOD), DIS (DISGEO), FRG (FRGMNT), HOM (HOMOLOGY de Biosym), INT (interactive), LOO (LOOK), MOD (MODELLER), SCR (utilisation de la chaîne principale de la structure parente la plus homologue), SCRave (utilisation d'une structure moyennée à partir de plusieurs structures parentes), SCRmix (utilisation de plusieurs fragments provenant de plusieurs protéines parentes en fonction de leurs homologies locales de séquences), STE (STEREO), WI (WHATIF).
Sixième colonne (GRAF) : principal outil graphique : ATOM (Alabama TOM), BI (BIOGRAPH), BRA (BRAGI), BRU (BRUGEL), CHE (CHEM), FR (FRODO), IN (INSIGHT de Biosym), MA (MANOSK), MEC (construction mécanique), MI (MIDAS), QU (QUANTA de MSI), SY (SYBYL), TF (TURBO-FRODO), XF (XFIT), WI (WHATIF).
Septième colonne (MIN) : champ de potentiels utilisé durant la minimisation : AM (AMBER), CH (CHARMM), CO (CONTACT), DI (DISCOVER), DM (DISMAN), DR (DREIDING), EC (ECEPP), GR (GROMOS), IM (IMPACT), LU (LUCIFER), MF (MODELFIT), PA (PAKGGRAF), PR (PRESTO), TR (TRIPOS), VF (VFFPRG), XP (XPLOR), WI (WHATIF).
Huitième colonne (RMS) : écart quadratique moyen entre le modèle et la structure parente. Lorsque que la structure cristallographique cible est disponible le RMS est indiqué en gras. Cα indique carbone α, bk inclue les atomes de la chaîne principale, core est le coeur de la protéine.
Neuvième colonne (DOCKING) : assemblage de substrats ou d'inhibiteurs dans le modèle. Le positionnement du substrat est obtenu soit de manière interactive (INT), soit transposé d'une structure de complexe connue (TRA). Le positionnement peut être affiné soit par minimisation (min) ou par dynamique moléculaire (MD).
Dixième colonne (AUTEURS) : réference.

REFERENCES

Abagyan R., Batalov S., Cardozo T., Totrov M., Webber J. & Zhou Y. (1997). Homology modeling with internal coordinate mechanics: Deformation zone mapping and improvements of models via conformational search. Proteins **Suppl. 1**, 29-37.

Abagyan R., Totrov M. & Kuznetsov D. (1994). ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorded native conformation. J. Comp. Chem. **15**, 488-506.

Abagyan R.A. & Batalov S. (1997). Do aligned sequences share the same fold? J. Mol. Biol. **273**, 355-368.

Adams P.D., Arkin I.T., Engelman D.M. & Brünger A.T. (1995). Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. Nature Struct. Biol. **2**, 154-162.

Adzhubei A.A., Laughton C.A. & Neidle S. (1995). An approach to protein homology modelling based on an ensemble of NMR structures: Application to the Sox-5 HMG-box protein. Prot. Eng. **8**, 615-625.

Aehle W., Sobek H., Amory A., Vetter R., Wilke D. & Schomburg D. (1993). Rational protein engineering and industrial application: Structure prediction by homology and rational design of protein-variants with improved 'washing performance'-the alkaline protease from *Bacillus alcalophilus*. J. Biotech. **28**, 31-40.

Aehle W., Sobek H. & Schomburg D. (1995). Evaluation of protein 3-D structure prediction: Comparison of modelled and X-ray structure of an alkaline serine protease. J. Biotech. **41**, 211-219.

Akahane K., Umeyama H., Nakagawa S., Moriguchi I., Hirose S., Iizuka K. & Murakami K. (1985). Three-dimensional structure of human renin. Hypertension **7**, 3-12.

Alberg D.G. & Schreiber S.L. (1993). Structure-based design of a cyclophilin-calcineurin bridging ligand. Science **262**, 248-250.

Alexandrov N.N. & Leuthy R. (1998). Alignment algorithm for homology modeling and threading. Protein Sci. **7**, 254-258.

Almassy R.J. & Dickerson R.E. (1978). Pseudomonas cytochrome c_{551} at 2.0 Å resolution: Enlargement of the cytochrome c family. Proc. Natl. Acad. Sci. USA **75**, 2674-2678.

Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**, 403-410.

Anfinsen C.B. (1973). Principles that govern the folding of protein chains. Science **181**, 223-230.

Annereau J.-P., Stoven V., Bontems F., Barthe J., Lenoir G., Blanquet S. & Lallemand J.-Y. (1997). Insight into cystic fibrosis by structural modelling of CFTR first nucleotide binding fold (NBF1). C. R. Acad. Sci. III-Vie **320**, 113-121.

Appelt K., Bacquet R.J., Barlett C.A., Booth C.L.J., Freer S.T., Fuhry M.A.M., Gehring M.R., Herrmann S.M., Howland E.F., Janson C.A., Jones T.R., Kan C.-C., Kathardekar V., Lewis K.K., Marzoni G.P., Matthews D.A., Mohr C., Moomaw E.W., Morse C.A., Oatley S.J., Ogden R.C., Reddy M.R., Reich S.H., Schoettlin W.S., Smith W.W., Varney M.D., Villafranca J.E., Ward R.W., Webber S., Webber S.E., Welsh K.M. & White J. (1991). Design of enzyme inhibitors using iterative protein crystallographic analysis. J. Med. Chem. **34**, 1925-1934.

Aszódi A. & Taylor W.R. (1996). Homology modelling by distance geometry. Fold. Design **1**, 325-334.

Bairoch A. (1991). PROSITE: A dictionary of sites and patterns in proteins. Nucleic Acids Res. **19**, 2241-2245.

Bajorath J., Stenkamp R. & Aruffo A. (1993). Knowledge-based model building of proteins: Concepts and examples. Protein Sci. **2**, 1798-1810.

Baldwin J.J., Ponticello G.S., Anderson P.S., Christy M.E., Murcko M.A., Randall W.C., Schwam H., Sug-

rue M.F., Springer J.P., Gautheron P., Grove J., Mallorga P., Viader M.-P., McKeever B.M. & Navia M.A. (1989). Thienothiopyran-2-sulfonamides: Novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma. J. Med. Chem. **32**, 2510-2513.

Bamborough P., Grant G.H., Hedgecock C.J.R., West S.P. & Richards W.G. (1993). A computer model of the interleukin-4/receptor complex. Proteins **17**, 11-19.

Bates P.A., Luo J. & Sternberg M.J.E. (1992). A predicted three-dimensional structure for the carcinoembryonic antigen (CEA). FEBS Lett. **301**, 207-214.

Bates P.A., McGregor M.J., Islam S.A., Sattentau Q.J. & Sternberg M.J.E. (1989). A predicted three-dimensional structure for the human immunodeficiency virus binding domains of CD4 antigen. Prot. Eng. **3**, 13-21.

Bax B., Blaber M., Ferguson G., Sternberg M.J.E. & Walls P.H. (1993). Prediction of the three-dimensional structures of the nerve growth factor and epidermal growth factor binding proteins (kallikreins) and an hypothetical structure of the high molecular weight complex of epidermal growth factor with its binding protein. Protein Sci. **2**, 1229-1241.

Bedarkar S., Blundell T.L., Gowan L.K., McDonald J.K. & Schwabe C. (1982). On the three-dimensional structure of relaxin. Ann. N.Y. Acad. Sci. **380**, 22-33.

Bedarkar S., Turnell W.G. & Blundell T.L. (1977). Relaxin has conformational homology with insulin. Nature **270**, 449-451.

Bell C.W., Roberts V.A., Scholthof K.-B.G., Zhang G. & Karu A.E. (1995). Recombinant antibodies to diuron: a model for the phenylurea combining site. In Immunoanalysis of agrochemicals. Emerging technologies (Nelson J.O., Karu A.E. & Wong R.B., eds.), pp. 50-71. American Chemical Society, Washington D.C.

Bhat T.N., Sasisekharan V. & Vijayan M. (1979). An analysis of side-chain conformation in proteins. Int. J. Peptide Protein Res. **13**, 170-184.

Bhowmick N., Huang J., Puett D., Isaacs N.W. & Lapthorn A.J. (1996). Determination of residues important in hormone binding to the extracellular domain of the luteinizing hormone/chorionic gonadotropin receptor by site-directed mutagenesis and modeling. Mol. Endocrinol. **10**, 1147-1159.

Bienstock R.J., Darden T., Wiseman R., Pedersen L. & Barrett J.C. (1996). Molecular modeling of the amino-terminal zinc ring domain of BRCA1. Cancer Res. **56**, 2539-2545.

Blaney J.M., Jorgensen E.C., Connolly M.L., Ferrin T.E., Langridge R., Oatley S.J., Burridge J.M. & Blake C.C.F. (1982). Computer graphics in drug design: Molecular modeling of thyroid hormone-prealbumin interactions. J. Med. Chem. **25**, 785-790.

Blundell T., Carney D., Gardner S., Hayes F., Howlin B., Hubbard T., Overington J., Singh D.A., Sibanda B.L. & Sutcliffe M. (1988). Knowledge-based protein modelling and design. Eur. J. Biochem. **172**, 513-520.

Blundell T. & Horuk R. (1981). A monomeric insulin from the casiragua: Molecular model building using computer graphics. Z. Physiol. Chem. **362**, 727-733.

Blundell T., Sibanda B.L. & Pearl L. (1983). Three-dimensional structure, specificity and catalytic mechanism of renin. Nature **304**, 273-275.

Blundell T.L., Bedarkar S., Rinderknecht E. & Humbel R.E. (1978). Insulin-like growth factor: A model for tertiary structure accounting for immunoreactivity and receptor binding. Proc. Natl. Acad. Sci. USA **75**, 180-184.

Blundell T.L., Sibanda B.L., Sternberg M.J.E. & Thornton J.M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. Nature **326**, 347-352.

Bower M.J., Cohen F.E. & Dunbrack R.L., Jr. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. J. Mol. Biol. **267**, 1268-1282.

Bowie J.U., Reidhaar-Olson J.F., Lim W.A. & Sauer R.T. (1990). Deciphering the message in protein sequences: Tolerance to amino acid substitutions. Science **247**, 1306-1310.

Bravi G., Legname G. & Chan A.W.E. (1995). Substrate recognition by ribosome-inactivating protein studied by molecular modeling and molecular electrostatic potentials. J. Mol. Graph. **13**, 83-88.

Brayer G.D. (1987). A preliminary structure for the DNA binding protein from bacteriophage IKe. J. Biomol. Struct. Dynam. **4**, 859-868.

Bridon D.P. & Dowell b.L. (1995). Structural comparison of prostate-specific antigen and human glandular kallikrein using molecular modeling. Urology **45**, 801-806.

Brocklehurst S.M. & Perham R.N. (1993). Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and the lipoylated H protein from the pea leaf glycine cleavage system: A new automated method for the prediction of protein tertiary structure. Protein Sci. **2**, 626-639.

Brooks B., Bruccoleri R., Olafson B., States D., Swaminathan S. & Karplus M. (1983). CHARMM: A program for macromolecular energy, minimization, and molecular dynamics calculations. J. Comp. Chem. **4**, 187-217.

Brown M., Rittenberg M.B., Chen C. & Roberts V.A. (1996). Tolerance to single, but not multiple, amino acid replacements in antibody VH CDRH2. A means of minimizing B cell wastage from somatic hypermutation? J. Immunol. **156**, 3285-3291.

Browne W.J., North A.C.T., Phillips D.C., Brew K., Vanaman T.C. & Hill R.L. (1969). A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. J. Mol. Biol. **42**, 65-86.

Bruccoleri R.E. (1993). Application of systematic conformational search to protein modeling. Mol. Simul. **10**, 151-174.

Bruccoleri R.E., Haber E. & Novotny J. (1988). Structure of antibody hypervariable loops reproduced by a conformational search algorithm. Nature **335**, 564-568.

Brünger A.T. (1992). X-PLOR Manual. Version 3.0, Yale University, New Haven.

Bruschi M., Bonicel J., Hatchikian E.C., Fardeau M.L., Belaich J.P. & Frey M. (1991). Amino acid sequence and molecular modelling of a thermostable two (4Fe-4S) ferredoxin from the archaebacterium *Methanococcus thermolithotrophicus*. Biochim. Biophys. Acta **1076**, 79-85.

Bryant S.H. & Amzel L.M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. Int. J. Peptide Protein Res. **29**, 46-52.

Cabral M.J.H., Lee A., Cohen S.L., Chait B.T., Li M. & Mackinnon R. (1998). Crystal structure and functional analysis of the HERG potassium channel N terminus: A eukaryotic PAS domain. Cell **95**, 649-655.

Cachau R.E., Erickson J.W. & Villar H.O. (1994). Novel procedure for structure refinement in homology modeling and its application to the human class Mu glutathione S-transferases. Prot. Eng. **7**, 831-839.

Caputo A., James M.N.G., Powers J.C., Hudig D. & Bleackley R.C. (1994). Conversion of the substrate specificity of mouse proteinase granzyme B. Nature Struct. Biol. **1**, 364-367.

Carlacci L. & Englander S.W. (1993). The loop problem in proteins: A Monte Carlo simulated annealing approach. Biopolymers **33**, 1271-1286.

Carlacci L. & Englander S.W. (1996). Loop problem in proteins: Developments on the Monte Carlo simulated annealing approach. J. Comp. Chem. **17**, 1002-1012.

Carlson W., Karplus M. & Haber E. (1985). Construction of a model for the three-dimensional structure of human renal renin. Hypertension **7**, 13-26.

Carson M., Bugg C.E., DeLucas L.J. & Narayana S.V.L. (1994). Comparison of homology models with the experimental structure of a novel serine protease. Acta Cryst. D **50**, 889-899.

Castonguay L.A., Bryant S.H., Snow P.M. & Fetrow J.S. (1995). A proposed structural model of domain 1 of fasciclin III neural cell adhesion protein based on an inverse folding algorithm. Protein Sci. **4**, 472-483.

Chandrasekaran R. & Ramachandran G.N. (1970). Studies on the conformation of amino acids. XI. Analysis of the observed side group conformations in proteins. Int. J. Protein Res. **2**, 223-233.

Chang Y.T., Stiffelman O.B. & Loew G.H. (1996). Computer modeling of 3D structures of cytochrome P450s. Biochimie **78**, 771-779.

Chang Y.-T. & Loew G.H. (1996). Construction and evaluation of a three-dimensional structure of cytochrome P450choP (CYP105C1). Prot. Eng. **9**, 755-766.

Chelvanayagam G., Roy G. & Argos P. (1994). Easy adaptation of protein structure to sequence. Prot. Eng. **7**, 173-184.

Chen C., Roberts V.A., Stevens S., Brown M., Stenzel-Poore M.P. & Rittenberg M.B. (1995). Enhancement and destruction of antibody function by somatic mutation: Unequal occurence is controlled by V gene combinatorial associations. EMBO J. **14**, 2784-2794.

Chen J.M., Grad R., Monaco R. & Pincus M.R. (1996a). Prediction of the three-dimensional structure of the rap-1A protein from its homology to the ras-gene-encoded p21 protein. J. Prot. Chem. **15**, 11-15.

Chen X., Whitmire D. & Bowen J.P. (1996b). Xylanase homology modeling using the inverse protein folding approach. Protein Sci. **5**, 705-708.

Chiche L., Gregoret L.M., Cohen F.E. & Kollman P. (1990). Protein model structure evaluation using the solvation free energy of folding. Proc. Natl. Acad. Sci. USA **87**, 3240-3243.

Chinea G., Padron G., Hooft R.W.W., Sander C. & Vriend G. (1995). The use of position-specific rotamers in model building by homology. Proteins **23**, 415-421.

Chothia C. (1975). Structural invariants in protein folding. Nature **254**, 304-308.

Chothia C. (1992). One thousand families for the molecular biologist. Nature **357**, 543-544.

Chothia C. & Lesk A.M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J. **5**, 823-826.

Chothia C. & Lesk A.M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. **196**, 901-917.

Chothia C., Lesk A.M., Tramontano A., Levitt M., Smith-Gill S.J., Air G., Sheriff S., Padlan E.A., Davies D., Tulip W.R., Colman P.M., Spinelli S., Alzari P.M. & Poljak R.J. (1989). Conformations of immunoglobulin hypervariable regions. Nature **342**, 877-883.

Chung S.Y. & Subbiah S. (1996). A structural explanation for the twilight zone of protein sequence homology. Structure **4**, 1123-1127.

Claessens M., van Cutsem E., Lasters I. & Wodak S. (1989). Modelling the polypeptide backbone with 'spare parts' from known protein structures. Prot. Eng. **2**, 335-345.

Clark M., Cramer R.D., III & van Opdenbosch N. (1989). Validation of the general purpose Tripos 5.2 force field. J. Comp. Chem. **10**, 982-1012.

Clement-Metral J.D., Holmgren A., Cambillau C., Jörnvall H., Eklund H., Thomas D. & Lederer F. (1988). Amino acid determination and three-dimensional modelling of thioredoxin from the photosynthetic bacterium *Rhodobacter sphaeroides* Y. Eur. J. Biochem. **172**, 413-419.

Cohen B.I., Presnell S.R. & Cohen F.E. (1993). Origins of structural diversity within sequentially identical hexapeptides. Protein Sci. **2**, 2134-2145.

Collura V., Higo J. & Garnier J. (1993). Modeling of protein loops by simulated annealing. Protein Sci. **2**, 1502-1510.

Coombs G.S., Bergstrom R.C., Pellequer J.L., Baker S.I., Navre M., Smith, M.M., Tainer J.A., Madison E.L. & Corey D.R. (1998). Substrate specificity of Prostate-specific antigen (PSA). Chem & Biol. **5**, 475-478.

Cox J.A., Alard P. & Schaad O. (1990). Comparative molecular modeling of Amphioxus calcium vector protein with calmodulin and troponin C. Prot. Eng. **4**, 23-32.

Cregut D., Liautard J.P., Heitz F. & Chiche L. (1993). Molecular modeling of coiled-coil α-tropomyosin:

Analysis of staggered and in register helix-helix interactions. Prot. Eng. **6**, 51-58.
Cregut D., Liautard J.-P. & Chiche L. (1994). Homology modelling of annexin I: Implicit solvation improves side-chain prediction and combination of evaluation criteria allows recognition of different types of conformational error. Prot. Eng. **7**, 1333-1344.
Cronet P., Sander C. & Vriend G. (1993). Modeling of transmembrane seven helix bundles. Prot. Eng. **6**, 59-64.
Curtis B.M., Presnell S.R., Srinivasan S., Sassenfeld H., Klinke R., Jeffery E., Cosman D., March C.J. & Cohen F.E. (1991). Experimental and theoretical studies of the three-dimensional structure of human interleukin-4. Proteins **11**, 111-119.
Dashper S.G., Huq K.N.L., Riley P.F. & Reynolds E.C. (1994). Complete amino acid sequence and comparative molecular modelling of Hpr from *Streptococcus mutans* ingbritt. Biochem. Biophys. Res. Commun. **199**, 1297-1304.
Dayhoff M.O. (1969). Atlas of protein sequence and structure, Maryland: National Biomedical research foundation, Silver Spring.
de Groot M.J., Vermeulen N.P.E., Mullenders D.L.J. & den Kelder G.M.D.-O. (1996). A homology model for rat Mu class glutathione S-transferase 4-4. Chem. Res. Toxicol. **9**, 28-40.
de la Paz P., Sutton B.J., Darsley M.J. & Rees A.R. (1986). Modelling of the combining sites of three anti-lysozyme monoclonal antibodies and of the complex between one of the antibodies and its epitope. EMBO J. **5**, 415-425.
De Maeyer M., Desmet J. & Lasters I. (1997). All in one: A highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. Folding & Design **2**, 53-66.
Defay T. & Cohen F.E. (1995). Evaluation of current techniques for *ab initio* protein structure prediction. Proteins **23**, 431-445.
Delbaere L.T.J., Brayer G.D. & James M.N.G. (1979). Comparison of the predicted model of α-lytic protease with the X-ray structure. Nature **279**, 165-168.
Desmet J., De Maeyer M., Hazes B. & Lasters I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. Nature **356**, 539-542.
Di Francesco V., Geetha V., Garnier J. & Munson P.J. (1997). Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. Proteins **Suppl. 1**, 123-128.
Dodson G.G., Eliopoulos E.E., Isaacs N.W., McCall M.J., Niall H.D. & North A.C.T. (1982). Rat relaxin: Insuline-like fold predicts a likely receptor binding region. Int. J. Biol. Macromol. **4**, 399-405.
Doolittle R.F. (1981). Similar amino acid sequences: Chance or common ancestry? Science **214**, 149-159.
Dorit R.L., Schoenbach L. & Gilbert W. (1990). How big is the universe of exons? Science **250**, 1377-1382.
Du P., Collins J.R. & Loew G.H. (1992). Homology modeling of a heme protein, lignin peroxidase, from the crystal structure of cytochrome c peroxidase. Prot. Eng. **5**, 679-691.
Dudek M.J. & Scheraga H.A. (1990). Protein structure prediction using a combination of sequence homology and global energy minimization. I. Global energy minimization of surface loops. J. Comp. Chem. **11**, 121-151.
Dunbrack R.L., Jr. & Cohen F.E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. **6**, 1661-1681.
Dunbrack R.L., Jr. & Karplus M. (1993). Backbone-dependent rotamer library for proteins application to side-chain prediction. J. Mol. Biol. **230**, 543-574.
Ealick S.E., Babu Y.S., Bugg C.E., Erion M.D., Guida W.C., Montgomery J.A. & Secrist III J.A. (1991). Application of crystallographic and modeling methods in the design of purine nucleoside phosphorylase inhibitors. Proc. Natl. Acad. Sci. USA **88**, 11540-11544.
Eisenhaber F., Persson B. & Argos P. (1995). Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino acid sequence. Crit. Rev. Biochem. Mol. Biol. **30**,

1-94.

Eisenmenger F., Argos P. & Abagyan R. (1993). A method to configure protein side-chains from the main-chain trace in homology modelling. J. Mol. Biol. **231**, 849-860.

Eklund H., Cambillau C., Sjöberg B.M., Holmgren A., Jörnvall H., Höög J.O. & Brändén C.I. (1984). Conformational and functional similarities between glutaredoxin and thioredoxins. EMBO J. **3**, 1443-1449.

Elsliger M.-A. & Potier M. (1994). Homologous modeling of the lysosomal protective protein/carboxypeptidase L: Structural and functional implications of mutations identified in galactosialidosis patients. Proteins **18**, 81-93.

Erion M.D., Niwas S., Rose J.D., Ananthan S., Allen M., Secrist III J.A., Babu Y.S., Bugg C.E., Guida W.C., Ealick S.E. & Montgomery J.A. (1993). Structure-based design of inhibitors of purine nucleoside phosphorylase. 3. 9-arylmethyl derivatives of 9-deazaguanine substituted on the methylene group. J. Med. Chem. **36**, 3771-3783.

Ermer O. (1976). Calculation of molecular properties. Struct. Bonding **27**, 163-211.

Eschenburg S., Krauspenhaar R., Mikhailov A., Stoeva S., Betzel C. & Voelter W. (1998). Primary structure and molecular modeling of mistletoe lectin I from *Viscum album*. Biochem. Biophys. Res. Commun. **247**, 367-372.

EU 3-D Validation Network. (1998). Who Checks the checkers? Four validation tools applied to eight atomic resolution structures. J. Mol. Biol. **276**, 417-436.

Fechteler T., Dengler U. & Schomburg D. (1995). Prediction of protein three-dimensional structures in insertion and deletion regions: A procedure for searching data bases of representative protein fragments using geometric scoring criteria. J. Mol. Biol. **253**, 114-131.

Feldmann R.J., Bing D.H., Furie B.C. & Furie B. (1978). Interactive computer surface graphics approach to study of the active site of bovine trypsin. Proc. Natl. Acad. Sci. USA **75**, 5409-5412.

Feldmann R.J., potter M. & Glaudemans C.P.J. (1981). A hypothetical space-filling model of the V-regions of the galactan-binding myeloma immunoglobulin J539. Molec. Immunol. **18**, 683-698.

Fernández J.A., Villoutreix B.O., Hackeng T.M., Griffin J.H. & Bouma B.N. (1994). Analysis of protein S C4b-binding protein interactions by homology modeling and inhibitory antibodies. Biochemistry **33**, 11073-11078.

Fersht A.R., Shi J.-P., Knill-Jones J., Lowe D.M., Wilkinson A.J., Blow D.M., Brick P., Carter P., Waye M.M.Y. & Winter G. (1985). Hydrogen bonding and biological specificity analysed by protein engineering. Nature **314**, 235-238.

Fidelis K., Stern P.S., Bacon D. & Moult J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. Prot. Eng. **7**, 953-960.

Fields B.A., Guss J.M. & Freeman H.C. (1991). Three-dimensional model for stellacyanin, a "blue" copper-protein. J. Mol. Biol. **222**, 1053-1065.

Filizola M., Perez J.J., Palomer A. & Mauleón D. (1997). Comparative molecular modeling study of the three-dimensional structures of prostaglandin endoperoxide H2 synthase 1 and 2 (COX-1 and COX-2). J. Mol. Graph. Model. **15**, 290-300.

Findlay J.B.C. & Pappin D.J.C. (1986). The opsin family of proteins. Biochem. J. **238**, 625-642.

Fine R.M., Wang H., Shenkin P.S., Yarmush D.L. & Levinthal C. (1986). Predicting antibody hypervariable loop conformations II: minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. Proteins **1**, 342-362.

Fischer D. & Eisenberg D. (1996). Fold recognition using sequence-derived predictions. Protein Sci. **5**, 947-955.

Fisher C.L., Greengard J.S. & Griffin J.H. (1994). Models of the serine protease domain of the human antithrombotic plasma factor activated protein C and its zymogen. Protein Sci. **3**, 588-599.

Fitch W.M. (1966). An improved method of testing for evolutionary homology. J. Mol. Biol. **16**, 9-16.

Flöckner H., Braxenthaler M., Lackner P., Jaritz M., Ortner M. & Sippl M.J. (1995). Progrss in fold recognition. Proteins **23**, 376-386.

Flöckner H., Domingues F.S. & Sippl M.J. (1997). Protein folds from pair interactions: A blind test in fold recognition. Proteins **Suppl. 1**, 129-133.

Floegel R., Zielenkiewicz P. & Saenger W. (1990). Tertiary structure of RNase Pch1 predicted from the model structure of RNase Ms and the crystal structure of RNase T1. Comparison among the model structures-testing the limits of modelling by homology. Eur. Biophys. J. **18**, 225-233.

Flores T.P., Orengo C.A., Moss D.S. & Thornton J.M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. **2**, 1811-1826.

Folkers G., Alber F., Amrhein I., Behrends H., Bohner T., Gerber S., Kuonen O. & Scapozza L. (1997). Integrated homology modelling and X-ray study of herpes simplex virus I thymidine kinase: A case study. J. Receptor Signal Transduc. Res. **17**, 475-494.

Frampton J., Gibson T.J., Ness S.A., Döderlein G. & Graf T. (1991). Proposed structure for the DNA-binding domain of the Myb oncoprotein based on model building and mutational analysis. Prot. Eng. **4**, 891-901.

Fujiyoshi-Yoneda T., Yoneda S., Kitamura K., Amisaki T., Ikeda K., Inoue M. & Ishida T. (1991). Adaptability of restrained molecular dynamics for tertiary structure prediction: Application to Crotalus atrox venom phospholipase A2. Prot. Eng. **4**, 443-450.

Furie B., Bing D.H., Feldmann R.J., Robison D.J., Burnier J.P. & Furie B.C. (1982). Computer-generated models of blood coagulation factor Xa, factor IXa, and thrombin based uopn structural homology with other serine proteases. J. Biol. Chem. **257**, 3875-3882.

Garnier J. (1990). Protein structure prediction. Biochimie **72**, 513-524.

Gelin B.R. & Karplus M. (1975). Side chain torsional potentials and motion of amino acids in proteins: Bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. USA **72**, 2002-2006.

Ghetti A., Bolognesi M., Cobianchi F. & Morandi C. (1990). Modelling by homology of RNA binding domain. Mol. Biol. Reports **14**, 87-88.

Gill D.S., Roush D.J. & Willson R.C. (1994). Tertiary structure of the heme0binding domian of rat cytochrome b_5 based on homology modeling. J. Biomol. Struct. Dynam. **11**, 1003-1015.

Giranda V.L., Chapman M.S. & Rossmann M.G. (1990). Modeling of the human intercellular adhesion molecule-1, the human rhinovirus major group receptor. Proteins **7**, 227-233.

Go N. & Scheraga H.A. (1970). Ring closure and local conformational deformation of chain molecules. Macromolecules **3**, 178-187.

Goldstein R.F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys. J. **66**, 1335-1340.

Gomar J., Sodano P., Ptak M. & Vovelle F. (1997). Homology modelling of an antimicrobial protein, Ace-AmP1, from lipid transfer protein structures. Folding & Design **2**, 183-192.

Gong W., Hao B., Mansy S.S., Gonzalez G., Gilles-Gonzales M.A. & Chan M.K. (1998) Structure of a biological oxygen sensor: A new mechanism for heme-driven signal transduction. Proc. Natl. Acad. Sci. USA **95**, 15177-15182.

Gonnet G.H., Cohen M.A. & Benner S.A. (1992). Exhaustive matching of the entire protein sequence database. Science **256**, 1443-1445.

Gready J.E., Ranganathan S., Schofield P.R., Matsuo Y. & Nishikawa K. (1997). Predicted structure of the extracellular region of ligand-gated ion-channel receptors shows SH2-like and SH3-like domains forming the ligand-binding site. Protein Sci. **6**, 983-998.

Greengard J.S., Fisher C.L., Villoutreix B. & Griffin J.H. (1994). Structural basis for type I and type II deficiencies of antithrombotic plasma protein C: Patterns revealed by three-dimensional molecular modeling of mutations of the protease domain. Proteins **18**, 367-380.

Greer J. (1980). Model for haptoglobin heavy chain based upon structural homology. Proc. Natl. Acad. Sci.

USA **77**, 3393-3397.

Greer J. (1981a). Comparative model-building of the mammalian serine proteases. J. Mol. Biol. **153**, 1027-1042.

Greer J. (1981b). Model of a specific interaction. Salt-bridges form between prothrombin and its activating enzyme blood clotting factor Xa. J. Mol. Biol. **153**, 1043-1053.

Greer J. (1990). Comparative modeling methods: Application to the family of the mammalian serine proteases. Proteins **7**, 317-334.

Greer J. (1991). Comparative modeling of homologous proteins. Meth. Enzymol. **202**, 239-252.

Gregory D.S., Martin A.C.R., Cheetham J.C. & Rees A.R. (1993). The prediction and characterization of metal binding sites in proteins. Prot. Eng. **6**, 29-35.

Gronenborn A.M. & Clore G.M. (1991). Modeling the three-dimensional structure of the monocyte chemoattractant and activating protein MCAF/MCP-1 on the basis of the solution structure of interleukin-8. Prot. Eng. **4**, 263-269.

Grossmann J.G., Ingleddew W.J., Harvey I., Strange R.W. & Hasnain S.S. (1995). X-ray absorption studies and homology modeling define the structural features that specify the nature of the copper site in rusticyanin. Biochemistry **34**, 8406-8414.

Grötzinger J., Engels M., Jacoby E., Wollmer A. & Straßburger W. (1991). A model for the C5a receptor and for its interaction with the ligand. Prot. Eng. **4**, 767-771.

Gunasekaran K., Ramakrishnan C. & Balaram P. (1996). Disallowed Ramachandran conformations of amino acid residues in protein structures. J. Mol. Biol. **254**, 191-198.

Gupta H.M., Talwar G.P. & Salunke D.M. (1993). A novel computer modeling approach to the structures of small bioactive peptides: The structure of Gonadotropin Releasing Hormone. Proteins **16**, 48-56.

Hagler A.T., Huler E. & Lifson S. (1974). Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. J. Amer. Chem. Soc. **96**, 5319-5327.

Han K.F. & Baker D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc. Natl. Acad. Sci. USA **93**, 5814-5818.

Hansch C., Li R.-L., Blaney J.M. & Langridge R. (1982). Comparison of the inhibition of Escherichia coli and Lactobacillus casei dihydrofolate reductase by 2,4-diamino-5-(substituted-benzyl)pyrimidines: Quantitative structure-activity relationships, X-ray crystallography, and computer graphics in structure-activity analysis. J. Med. Chem. **25**, 777-784.

Hartley B.S. (1970). Homologies in serine proteinases. Phil. Trans. Roy. Soc. Lond. B **257**, 77-87.

Havel T.F., Crippen G.M. & Kuntz I.D. (1979). Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. Biopolymers **18**, 73-81.

Havel T.F. & Wüthrich K. (1985). An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. J. Mol. Biol. **182**, 281-294.

Havel T.F. & Snow M.E. (1991). A new method for building protein conformations from sequence alignments with homologues of known structure. J. Mol. Biol. **217**, 1-7.

Havel T.F. (1993). Predicting the structure of the flavodoxin from Escherichia coli by homology modeling, distance geometry and molecular dynamics. Mol. Simul. **10**, 175-210.

Hearst D.P. & Cohen F.E. (1994). GRAFTER: A computational aid for the design of novel proteins. Prot. Eng. **7**, 1411-1421.

Heinz D.W., Baase W.A., Dahlquist F.W. & Matthews B.W. (1993). How amino-acid insertions are allowed in an α-helix of T4 lysozyme. Nature **361**, 561-564.

Hellinga H.W. & Richards F.M. (1991). Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. J. Mol. Biol. **222**, 763-785.

Hellinga H.W., Wynn R. & Richards F.M. (1992). The hydrophobic core of *Escherichia coli* thioredoxin shows a high tolerance to nonconservative single amino acid substitutions. Biochemistry **31**, 11203-11209.

Hendrickson J.B. (1961). Molecular geometry. I. Machine computation of the common rings. J. Amer. Chem. Soc. **83**, 4537-4547.

Henikoff S. & Henikoff J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**, 10915-10919.

Henrick K., Bawumia S., Barboni E.A.M., Mehul B. & Hughes R.C. (1998). Evidence for subsites in the galectins involved in sugar binding at the nonreducing end of the central galactose of oligosaccharide ligands: Sequence analysis, homology modeling and mutagenesis studies of hamster galectin-3. Glycobiology **8**, 45-57.

Henriques E.F., Ramos M.J. & Reynolds C.A. (1997). Inclusion of conserved buried water molecules in the model structure of rat submaxillary kallikrein. J. Comput-Aided Mol. Des. **11**, 547-556.

Higo J., Collura V. & Garnier J. (1992). Development of an extended simulated annealing method: Application to the modeling of complementary determining regions of immunoglobulins. Biopolymers **32**, 33-43.

Hilbert M., Böhm G. & Jaenicke R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. Proteins **17**, 138-151.

Hilpert K., Ackermann J., Banner D.W., Gast A., Gubernator K., Hadváry P., Labler L., Müller K., Schmid G., Tschopp T.B. & van de Waterbeemd H. (1994). Design and synthesis of potent and highly selective thrombin inhibitors. J. Med. Chem. **37**, 3889-3901.

Hoffrén A.M., Saloheimo M., Thomas P., Overington J., Johnson M.S. & Blundell T.L. (1991). Modelling the lignin peroxidase LIII of *Phlebia radiata* using a knowledge-based approach. J. Chim. Phys. **88**, 2659-2662.

Hoffrén A.-M., Saloheimo M., Thomas P., Overington J.P., Johnson M.S., Knowles J.K.C. & Blundell T.L. (1993). Modelling of the lignin peroxidase LIII of *Phlebia radiata*: Use of a sequence template generated from a 3-D structure. Prot. Eng. **6**, 177-182.

Hofmann K., Bucher P. & Kajava A.V. (1998). A model of Cdc25 phosphatase catalytic domain and Cdk-interaction surface based on the presence of a rhodanese homology domain. J. Mol. Biol. **282**, 195-208.

Holm L. & Sander C. (1997). New structure - novel fold? Structure **5**, 165-171.

Hornischer K. & Blöcker H. (1996). Grafting of discontinuous sites: A protein modeling strategy. Prot. Eng. **9**, 931-939.

Hu X., Xu D., Hamer K., Schulten K., Koepke J. & Michel H. (1995). Predicting the structure of the light-harvesting complex II of *Rhodospirillum molischianum*. Protein Sci. **4**, 1670-1682.

Hubbard T.J. & Park J. (1995). Fold recognition and *ab initio* structure predictions using hidden Markov models and β-strand pair potentials. Proteins **23**, 398-402.

Hubbard T.J.P. & Blundell T.L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. Prot. Eng. **1**, 159-171.

Hutschinson E.G. & Thornton J.M. (1996). PROMOTIF- A program to identify and analyze structural motifs in proteins. Protein Sci. **5**, 212-220.

Ibdah M., Bar-Ilan A., Livnah O., Schloss J.V., Barak Z. & Chipman D.M. (1996). Homology modeling of the structure of bacterial acetohydroxy acid synthase and examination of the active site by site-directed mutagenesis. Biochemistry **35**, 16282-16291.

Inana G., Piatigorsky J., Norman B., Slingsby C. & Blundell T. (1983). Gene and protein structure of a β-crystallin polypeptide in murine lens: Relationship of exons and structural motifs. Nature **302**, 310-315.

Isaacs N., James R., Niall H., Bryant-Greenwood G., Dodson G., Evans A. & North A.C.T. (1978). Relaxin

and its structural relationship to insulin. Nature **271**, 278-281.

Jäger J., Solmajer T. & Jansonius J.N. (1992). Computational approach towards the three-dimensional structure of *E. coli* tyrosine aminotransferase. FEBS Lett. **306**, 234-238.

Janin J., Wodak S., Levitt M. & Maigret B. (1978). Conformation of amino acid side-chains in proteins. J. Mol. Biol. **125**, 357-386.

Jarvis J.A., Munro S.L.A. & Craik D.J. (1992). Homology model of thyroxine binding globulin and elucidation of the thyroid hormone binding site. Prot. Eng. **5**, 61-67.

Johnson M.S. & Overington J.P. (1993). A structural basis for sequence comparisons. An evaluation of scoring methodologies. J. Mol. Biol. **233**, 716-738.

Johnson M.S., Srinivasan N., Sowdhamini R. & Blundell T.L. (1994). Knowledge-based protein modeling. Crit. Rev. Biochem. Mol. Biol. **29**, 1-68.

Jones D.T., Miller R.T. & Thornton J.M. (1995). Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins **23**, 387-397.

Jones D.T., Taylor W.R. & Thornton J.M. (1992). A new approach to protein fold recognition. Nature **358**, 86-89.

Jones T.A. (1978). A graphic model building and refinement system for macromolecules. J. Appl. Cryst. **11**, 258-272.

Jones T.A. & Thirup S. (1986). Using known substructures in protein model building and crystallography. EMBO J. **5**, 819-822.

Jorgensen W.L. & Tirado-Rives J. (1988). The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. J. Amer. Chem. Soc. **110**, 1657-1666.

Kabat E.A., Wu T.T. & Bilofsky H. (1977). Unusual distributions of amino acids in complementarity-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. J. Biol. Chem. **252**, 6609-6616.

Kabsch W. & Sander C. (1984). On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. USA **81**, 1075-1078.

Kajava A.V. (1998). Structural diversity of leucine-rich repeat proteins. J. Mol. Biol. **277**, 519-527.

Kajihara A., Komooka H., Kamiya K. & Umeyama H. (1993). Protein modelling using the chimera reference protein derived from exons. Prot. Eng. **6**, 615-620.

Katz D.S. & Christianson D.W. (1993). Modeling the uncleaved serpin antichymotrypsin and its chymotrypsin complex. Prot. Eng. **6**, 701-709.

Kauzmann W. (1959). Some factors in the interpretation of protein denaturation. Adv. Prot. Chem. **14**, 1-63.

Keen J.N., Caceres I., Eliopoulos E.E., Zagalsky P.F. & Findlay J.B.C. (1991). Complete sequence and model for the C1 subunit of the carotenoprotein, crustacyanin, and model for the dimer, b-crustacyanin, formed from the C1 and A2 subunits with astaxanthin. Eur. J. Biochem. **202**, 31-40.

Keresztessy Z. & Hughes M.A. (1998). Homology modelling and molecular dynamics aided analysis of ligand complexes demonstrates functional properties of lipid-transfer proteins encoded by the barley low-temperature-inducible gene family, blt4. Plant J. **14**, 523-533.

Kettleborough C.A., Saldanha J., Heath V.J., Morrison C.J. & Bendig M.M. (1991). Humanization of a mouse monoclonal antibody by CDR-grafting: The importance of framework residues on loop conformation. Prot. Eng. **4**, 773-783.

Khoti H., McLeod A.N., Blundell T.L., Ishizaki H., Nagasawa H. & Suzuki A. (1987). Prothoracicotropic hormone has an insulin-like structure. FEBS Lett. **219**, 419-425.

Kleywegt G.J. (1997). Validation of protein models from Cα coordinates alone. J. Mol. Biol. **273**, 371-376.

Kocher J.-P.A., Rooman M.J. & Wodak S.J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J. Mol. Biol. **235**, 1598-1613.

Koehl P. & Delarue M. (1995). A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modeling. Nature Struct. Biol. **2**, 163-170.

Kort R., Phillips-Jones M.K., van Aalten D.M.F., Haker A., Hoffer S.M., Hellingwerf K.J. & Crielaard W. (1998). Sequence, chromophore extraction and 3-D model of the photoactive yellow protein from phrodobacter sphaeroides. Biochim. Biophys. Acta **1385**, 1-6.

Kraulis P.J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 24, 946-950

Kretsinger R.H. & Barry C.D. (1975). The predicted structure of the calcium-binding component of troponin. Biochim. Biophys. Acta **405**, 40-52.

Kumar V.D. & Weber I.T. (1992). Molecular model of the cyclic GMP-binding domain of the cyclic GMP-gated ion channel. Biochemistry **31**, 4643-4649.

Kuyper L.F., Roth B., Baccanari D.P., Ferone R., Beddell C.R., Champness J.N., Stammers D.K., Dann J.G., Norrington F.E., Baker D.J. & Goodford P.J. (1985). Receptor-based design of dihydrofolate reductase inhibitors: Comparison of crystallographically determined enzyme binding with enzyme affinity in a series of carboxy-susbtituted trimethoprim analogues. J. Med. Chem. **28**, 303-311.

Lacroix M., Rossi V., Gaboriaud C., Chevallier S., Jaquinod M., Thielens N.M., Gagnon J. & Arlaud G.J. (1997). Structure and assembly of the catalytic region of human complement protease C1r: A three-dimensional model based on chemical cross-linking and homology modeling. Biochemistry **36**, 6270-6282.

Lam P.Y.S., Jadhav P.K., Eyermann C.J., Hodge C.N., Ru Y., Bacheler L.T., Meek J.L., Otto M.J., Rayner M.M., Wong Y.N., Chang C.-H., Weber P.C., Jackson D.A., Sharpe T.R. & Erickson-Viitanen S. (1994). Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. Science **263**, 380-384.

Langridge R., Ferrin T.E., Kuntz I.D. & Connolly M.L. (1981). Real-time color graphics in studies of molecular interactions. Science **211**, 661-666.

Lapadat M.A., Deerfield D.W., II, Pedersen L.G. & Spremulli L.L. (1990). Generation of potential structures for the G-domain of chloroplast EF-Tu using comparative molecular modeling. Proteins **8**, 237-250.

Lapatto R. (1991). Model for the structure of formaldehyde dehydrogenase based on alcohol dehydrogenase. Int. J. Biol. Macromol. **13**, 73-76.

Laskowski R.A., MacArthur M.W., Moss D.S. & Thornton J.M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. J. Appl. Cryst. **26**, 283-291.

Lasters I., Desmet J. & De Maeyer M. (1997). Dead-end based modeling tools to explore the sequence space that is compatible with a given scaffold. J. Prot. Chem. **16**, 449-452.

Laughton C.A. & Neidle S. (1990). A molecular model for the enzyme cytochrome $P450_{17a}$, a major target for the chemotherapy of prostatic cancer. Biochem. Biophys. Res. Commun. **171**, 1160-1167.

Laughton C.A. (1994a). Prediction of protein side-chain conformations from local three-dimensional homology relationships. J. Mol. Biol. **235**, 1088-1097.

Laughton C.A. (1994b). A study of simulated annealing protocols for use with molecular dynamics in protein structure prediction. Prot. Eng. **7**, 235-241.

Lee C. & Subbiah S. (1991). Prediction of protein side-chain conformation by packing optimization. J. Mol. Biol. **217**, 373-388.

Lee R. H. (1992) Protein model building using structural homology. Nature **356**, 543-544.

Lesk A.M. & Chothia C. (1980). How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. J. Mol. Biol. **136**, 225-270.

Lesser G.J. & Rose G.D. (1990). Hydrophobicity of amino acid subgroups in proteins. Proteins **8**, 6-13.

Levitt M. (1992). Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. **226**, 507-533.

Lewis D.F. & Lee-Robichaud P. (1998). Molecular modelling of steroidogenic cytochromes P450 from Families CYP11, CYP17, CYP19 and CYP21 based on the CYP102 crystal structure. J. Steroid Biochem. Mol. Biol. **66**, 217-233.

Lewis D.F.V., Eddershaw P.J., Goldfarb P.S. & Tarbit M.H. (1997). Molecular modeling of cytochrome P4502D6 (CYP2D6) based on an alignment with CYP102: Structural studies on specific CYP2D6 substrate metabolism. Xenobiotica **27**, 319-340.

Lewis D.F.V., Parker M.G. & King R.J.B. (1995). Molecular modelling of the human estrogen receptor and ligand interactions based on site-directed mutagenesis and amino acid sequence homology. J. Steroid Biochem. Mol. Biol. **52**, 55-65.

Li H., Tejero R., Monleon D., Bassolino-Klimas D., Abate-Shen C., Bruccoleri R.E. & Montelione G.T. (1997). Homology modeling using simulated annealing of restrained molecular dynamics and conformational search calculations with CONGEN: Application in predicting the three-dimensional structure of murine homeodomain Msx-1. Protein Sci. **6**, 956-970.

Lipke P.N., Chen M.-H., de Nobel H., Kurjan J. & Kahn P.C. (1995). Homology modeling of an immunoglobulin-like domain in the *Saccharomyces cerevisiae* adhesion protein α-agglutinin. Protein Sci. **4**, 2168-2178.

Lobaccaro J.M., Poujol N., Chiche L., Lumbroso S., Brown T.R. & Sultan C. (1996). Molecular modeling and in vitro investigations of the human androgen receptor DNA-binding domain: Application for the study of two mutations. Mol. Cell. Endocrinol. **116**, 137-147.

Lozano J.J., López-de-Brinas E., Centeno N.B., Guigó R. & Sanz F. (1997). Three-dimensional modelling of human cytochrome P450 1A2 and its interaction with caffeine and MeIQ. J. Comput.-Aided Mol. Design **11**, 395-408.

Luchin S.V., Zinovieva R.D., Tomarev S.I., Dolgilevich S.M., Gause G.G., Bax B., Jr., Driessen H. & Blundell T.L. (1987). Frog lens βA1-crystallin: The nucleotide sequence of the cloned cDNA and computer graphics modelling of the three-dimensional structure. Biochim. Biophys. Acta **916**, 163-171.

Lüthy R., Bowie J.U. & Eisenberg D. (1992). Assessment of protein models with three-dimensional profiles. Nature **356**, 83-85.

MacArthur M.W. & Thornton J.M. (1996). Deviations from planarity of the peptide bond in peptides and proteins. J. Mol. Biol. **264**, 1180-1195.

Mainhart C.R., Potter M. & Feldmann R.J. (1984). A refined model for the variable domains (Fv) of the J539 ß(1,6)-D-Galactan-binding immunoglobulin. Molec. Immunol. **21**, 469-478.

Martin A.C.R., Cheetham J.C. & Rees A.R. (1989). Modeling antibody hypervariable loops: A combined algorithm. Proc. Natl. Acad. Sci. USA **86**, 9268-9272.

Martin A.C.R., Cheetham J.C. & Rees A.R. (1991). Molecular modeling of antibody combining sites. Meth. Enzymol. **203**, 121-153.

Martin A.C.R. & Thornton J.M. (1996). Structural families in loops of homologous proteins: Automatic classification, modelling and application to antibodies. J. Mol. Biol. **263**, 800-815.

Martin A.C.R., MacArthur M.W. & Thornton J.M. (1997). Assessment of comparative modeling in CASP2. Proteins **Suppl. 1**, 14-28.

Mas M.T., Smith K.C., Yarmush D.L., Aisaka K. & Fine R.M. (1992). Modeling the anti-CEA antibody combining site by homology and conformational search. Proteins **14**, 483-498.

Mathiowetz A.M. & Goddard W.A., III. (1995). Building proteins from Cα coordinates using the dihedral probability grid Monte Carlo method. Protein Sci. **4**, 1217-1232.

Matsumoto R., Sali A., Ghildyal N., Karplus M. & Stevens R.L. (1995). Packaging of proteases and proteoglycans in the granules of mast cells and other hemetopoietic cells. A cluster of histidines on mouse mast cell protease 7 regulates its binding to heparin serglycin proteoglycans. J. Biol. Chem. **270**, 19524-19531.

Matthews D.A., Bolin J.T., Burridge J.M., Filman D.J., Volz K.W. & Kraut J. (1985). Dihydrofolate reduc-

tase. The stereochemistry of inhibitor selectivity. J. Biol. Chem. **260**, 392-399.
Mayo S.L., Olafson B. & Goddard W.A., III. (1990). DREIDING: A generic force field for molecular simulations. J. Phys. Chem. **94**, 8897-8909.
McLachlan A.D. (1971). Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c551. J. Mol. Biol. **61**, 409-424.
McLachlan A.D. & Shotton D.M. (1971). Structural similarities between α-lytic protease of Myxobacter 495 and elastase. Nature - New Biol. **229**, 202-205.
McLachlan A.D. (1972). A mathematical procedure for superimposing atomic coordinates of proteins. Acta Cryst. **A28**, 656-657.
Merritt E.A. & Bacon D.J. (1997). Raster3D: Photorealistic molecular graphics. Meth. Enzymol. **277**, 505-524.
Miller S., Janin J., Lesk A.M. & Chothia C. (1987). Interior and surface of monomeric proteins. J. Mol. Biol. **196**, 641-656.
Mills A. (1993). Modelling the carbohydrate recognition domain of human E-selectin. FEBS Lett. **319**, 5-11.
Mimura C.S., Holbrook S.R. & Ames G.F.-L. (1991). Structural model of the nucleotide-binding conserved component of periplasmic permeases. Proc. Natl. Acad. Sci. USA **88**, 84-88.
Minor D.L., Jr. & Kim P.S. (1996). Context-dependent secondary structure formation of a designed protein sequence. Nature **380**, 730-734.
Momany F.A., McGuire R.F., Burgess A.W. & Scheraga H.A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. J. Phys. Chem. **79**, 2361-2381.
Montgomery J.A., Niwas S., Rose J.D., Secrist III J.A., Babu Y.S., Bugg C.E., Erion M.D., Guida W.C. & Ealick S.E. (1993). Structure-based design of inhibitors of purine nucleoside phosphorylase. 1. 9-(arylmethyl) derivatives of 9-deazaguanine. J. Med. Chem. **36**, 55-69.
Morea V., Tramontano A., Rustici M., Chothia C. & Lesk A.M. (1998). Conformations of the third hypervariable region in the VH domain of immunoglobulins. J. Mol. Biol. **275**, 269-294.
Morris A.L., MacArthur M.W., Hutchinson E.G. & Thornton J.M. (1992). Stereochemical quality of protein structure coordinates. Proteins **12**, 345-364.
Mosimann S.C., Johns K.L., Ardelt W., Mikulski S.M., Shogen K. & James M.N.G. (1992). Comparative molecular modeling and crystallization of P-30 protein: A novel antitumor protein of *Rana pipiens* oocytes and early embryos. Proteins **14**, 392-400.
Mosimann S., Meleshko R. & James M.N.G. (1995). A critical assessment of comparative molecular modeling of tertiary structures of proteins. Proteins **23**, 301-317.
Moult J. & James M.N.G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins **1**, 146-163.
Mukhopadhyay C. & Lohia A. (1998). Modeling of *Entamoeba histolytica* ferredoxin. J. Biomol. Struct. Dynam. **15**, 663-672.
Mumenthaler C., Schneider U., Buchholz C.J., Koller D., Braun W. & Cattaneo R. (1997). A 3D model for the measles virus receptor CD46 based on homology modeling, Monte Carlo simulations, and hemagglutinin binding studies. Protein Sci. **6**, 588-597.
Murgolo N.J., Windsor W.T., Hruza A., Reichert P., Tsarbopoulos A., Baldwin S., Huang E., Pramanik B., Ealick S. & Trotta P.P. (1993). A homology model of human interferon α-2. Proteins **17**, 62-74.
Musat G.V., Neumann J.M., Smith J.C. & Sanson A. (1997). Structure of human annexin I: Comparison of homology modelling and crystallographic experiment. Biochimie **79**, 691-703.
Nakamura H., Katayanagi K., Morikawa K. & Ikehara M. (1991). Structural models of ribonuclease H domains in reverse transcriptases from retroviruses. Nucleic Acids Res. **19**, 1817-1823.

Nauss J.L., Reid R.H. & Sadegh-Nasseri S. (1995). Accuracy of a structural homology model for a class II histocompatibility protein, HLA-DR1: Comparison to the crystal structure. J. Biomol. Struct. Dynam. **12**, 1213-1233.

Needleman S.B. & Wunsch C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**, 443-453.

Neidle S. & Goodwin G.H. (1994). A homology-based molecular model of the proline-rich homeodomain protein Prh, from haematopoietic cells. FEBS Lett. **345**, 93-98.

Neil K.J., Ridsdale R.A., Rutherford B., Taylor L., Larson D.E., Glibetic M., Rothblum L.I. & Harauz G. (1996). Structure of recombinant rat UBF by electron image analysis and homology modelling. Nucleic Acids Res. **24**, 1472-1480.

Neumüller M. & Jähnig F. (1996). Modeling of halorhodopsin and rhodopsin based on bacteriorhodopsin. Proteins **26**, 146-156.

Nieba L., Honegger A., Krebber C. & Plückthun A. (1997). Disrupting the hydrophobic patches at the antibody variable/constant domain interface: Improved in vivo folding and physical characterization of an engineered scFv fragment. Prot. Eng. **10**, 435-444.

Niefind K. & Schomburg D. (1991). Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. J. Mol. Biol. **219**, 481-497.

Nishikawa K. & Ooi T. (1980). Prediction of the surface-interior diagram of globular proteins by an empirical method. Int. J. Peptide Protein Res. **16**, 19-32.

Nordvall G. & Hacksell U. (1993). Binding-site modeling of the muscarinic m1 receptor: A combination of homology-based and indirect approaches. J. Med. Chem. **36**, 967-976.

Novotny J., Rashin A.A. & Bruccoleri R.E. (1988). Criteria that disriminate between native proteins and incorrectly folded models. Proteins **4**, 19-30.

Ogata K. & Umeyana H. (1997). Prediction of protein side-chain conformations by principal component analysis for fixed main-chain atoms. Prot. Eng. **10**, 353-359.

Ogata K. & Umeyama H. (1998). The role played by environmental residues on sidechain torsional angles within homologous families of proteins: A new method of sidechain modeling. Proteins **31**, 355-369.

Ohlendorf D.H. (1994). Accuracy of refined protein structures. II. Comparison of four independently refined models of human interleukin 1β. Acta Cryst. D **50**, 808-812.

Oliva B., Bates P.A., Querol E., Avilés F.X. & Sternberg M.J.E. (1998). Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein prediction. J. Mol. Biol. **279**, 1193-1210.

O'Neil K.T. & DeGrado W.F. (1985). A predicted structure of calmodulin suggests an electrostatic basis for its funtion. Proc. Natl. Acad. Sci. USA **82**, 4954-4958.

Orengo C.A., Jones D.T. & Thornton J.M. (1994). Protein superfamilies and domain superfolds. Nature **372**, 631-634.

Pabo C.O. & Suchanek E.G. (1986). Computer-aided model-building strategies for protein design. Biochemistry **25**, 5987-5991.

Pace C.N., Shirley B.A., McNutt M. & Gajiwala K. (1996). Forces contributing to the conformational stability of protein. FASEB J. **10**, 75-83.

Padlan E.A., Davies D.R., Pecht I., Givol D. & Wright C. (1976). Model-building studies of antigen-binding sites: The hapten-binding site of MOPC-315. Cold Spring Harbor Symp. Quant. Biol. **XLI**, 627-637.

Padlan E.A. & Kabat E.A. (1988). Model-building study of the combining sites of two antibodies to α(1->6)dextran. Proc. Natl. Acad. Sci. USA **85**, 6885-6889.

Padlan E.A. & Helm B.A. (1993). Modeling of the lectin-homology domains of the human and murine low-affinity Fcε receptor (FcεRII/CD23). Receptor **3**, 325-341.

Palmer K.A., Scheraga H.A., Riordan J.F. & Vallee B.L. (1986). A preliminary three-dimensional struc-

ture of angiogenin. Proc. Natl. Acad. Sci. USA **83**, 1965-1969.
Pan Y., DeFay T., Gitschier J. & Cohen F.E. (1995). Proposed structure of the A domains of factor VIII by homology modelling. Nature Struct. Biol. **2**, 740-744.
Park B.H. & Levitt M. (1996). Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J. Mol. Biol. **258**, 367-392.
Park B.H., Huang E.S. & Levitt M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J. Mol. Biol. **266**, 831-846.
Pascarella S. & Argos P. (1992). Analysis of insertions/deletions in protein structures. J. Mol. Biol. **224**, 461-471.
Pawlowski K., Bierzynski A. & Godzik A. (1996). Structural diversity in a family of homologous proteins. J. Mol. Biol. **258**, 349-366.
Pearl L.H. & Taylor W.R. (1987). A structural model for the retroviral proteases. Nature **329**, 351-354.
Pellequer J.L., Chen S.-w.W., Tainer J.A. & Getzoff E.D. (1996). Analysis of small conformational changes in proteins. Protein Sci. **5, sup1**, 99.
Pellequer J.-L. & Chen S.-w.W. (1997). Does conformational free energy distinguish loop conformations in proteins? Biophys. J. **73**, 2359-2375.
Pellequer J.-L., Wager-Smith K.A., Kay S.A. & Getzoff E.D. (1998a). Photoactive yellow protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily. Proc. Natl. Acad. Sci. USA **95**, 5884-5890.
Pellequer J.L., Gale A.J., Griffin J.H. & Getzoff E.D. (1998b). Homology modeling of Factor Va, a cofactor of the prothrombinase complex. Protein Sci. **7,sup1**, 159.
Pellequer J.L., Gale A.J., Griffin J.H. & Getzoff E.D. (1998c). Homology models of the C domains of blood coagulation Factors V and VIII: A proposed membrane binding mode for FV and FVIII C2 domains. Blood Cells, Mol. Diseases. **24**, 448-461.
Pemberton S., Lindley P., Zaitsev V., Card G., Tuddenham E.G.D. & Kemball-Cook G. (1997). A molecular model for the triplicated A domains of human factor VIII based on the crystal structure of human ceruloplasmin. Blood **89**, 2413-2421.
Perentesis J.P., Phan L.D., Gleason W.B., LaPorte D.C., Livingston D.M. & Bodley J.W. (1992). *Saccharomyces cerevisiae* elongation factor 2. Genetic cloning, characterization of expression, and G-domain modeling. J. Biol. Chem. **267**, 1190-1197.
Perutz M.F., Kendrew J.C. & Watson H.C. (1965). Structure and function of haemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. J. Mol. Biol. **13**, 669-678.
Pisabarro M.T., Ortiz A.R., Serrano L. & Wade R.C. (1994). Homology modeling of the Ab1-SH3 domain. Proteins **20**, 203-215.
Plattner J.J., Greer J., Fung A.K.L., Stein H., Kleinert H.D., Sham H.L., Smital J.R. & Perun T.J. (1986). Peptide analogues of angiotensinogen effect of peptide chain length on renin inhibition. Biochem. Biophys. Res. Commun. **139**, 982-990.
Ponder J.W. & 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.
Poteete A.R., Rennell D. & Bouvier S.E. (1992). Functional significance of conserved amino acid residues. Proteins **13**, 38-40.
Presta L.G., Lahr S.J., Shields R.L., Porter J.P., Gorman C.M., Fendly B.M. & Jardieu P.M. (1993). Humanization of an antibody directed against IgE. J. Immunol. **151**, 2623-2632.
Quondam M., Barbato C., Pickford A., Helmer-Citterich M. & Macino G. (1997). Homology modeling of *Neurospora crassa* geranylgeranyl pyrophosphate synthase: Structural interpretation of mutant phenotypes. Prot. Eng. **10**, 1047-1055.
Rao A.G., Hassan M. & Hempel J.C. (1994). Structure-function validation of high lysine analogs of α-hordothionin designed by protein modeling. Prot. Eng. **7**, 1485-1493.

Reczko M., Martin A.C.R., Bohr H. & Suhai S. (1995). Prediction of hypervariable CDR-H3 loop structures in antibodies. Prot. Eng. **8**, 389-395.

Rennell D., Bouvier S.E., Hardy L.W. & Poteete A.R. (1991). Systematic mutation of bacteriophage T4 lysozyme. J. Mol. Biol. **222**, 67-87.

Renouf D. & Hounsell E.F. (1995). Molecular modelling of glycoproteins by homology with non-glycosylated protein domains, computer simulated glycosylation and molecular dynamics. Adv. Exp. Med. Biol. **376**, 37-45.

Rey A. & Skolnick J. (1992). Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. J. Comp. Chem. **13**, 443-456.

Rice D.W. & Eisenberg D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J. Mol. Biol. **267**, 1026-1038.

Riechmann L., Weill M. & Cavanagh J. (1992). Improving the antigen affinity of an antibody Fv-fragment by protein design. J. Mol. Biol. **224**, 913-918.

Ring C.S. & Cohen F.E. (1994). Conformational sampling of loop structure using genetic algorithms. Israel J. Chem. **34**, 245-252.

Ring C.S., Sun E., McKerrow J.H., Lee G.K., Rosenthal P.J., Kuntz I.D. & Cohen F.E. (1993). Structure-based inhibitor design by using protein models for the development of antiparasitic agents. Proc. Natl. Acad. Sci. USA **90**, 3583-3587.

Rippmann F., Taylor W.R., Rothbard J.B. & Green N.M. (1991). A hypothetical model for the peptide binding domain hsp70 based on the peptide binding domain of HLA. EMBO J. **10**, 1053-1059.

Risler J.L., Delorme M.O., Delacroix H. & Henaut A. (1988). Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J. Mol. Biol. **204**, 1019-1029.

Roberts V.A., Stewart J., Benkovic S.J. & Getzoff E.D. (1994). Catalytic antibody model and mutagenesis implicate arginine in transition-state stabilization. J. Mol. Biol. **235**, 1098-1116.

Robson B. & Platt E. (1987). Modelling of α-lactalbumin from the known structure of hen egg white lysozyme using molecular dynamics. J. Comput.-Aided Mol. Design **1**, 17-22.

Robson B., Platt E., Fishleigh R.V., Marsden A. & Millard P. (1987). Expert system for protein engineering: Its application in the study of chloramphenicol acetyltransferase and avian pancreatic polypeptide. J. Mol. Graph. **5**, 8-17.

Rosenbach D. & Rosenfeld R. (1995). Simultaneous modeling of multiple loops in proteins. Protein Sci. **4**, 496-505.

Rossi V., Gaboriaud C., Lacroix M., Ulrich J., Fontecilla-Camps J.C., Gagnon J. & Arlaud G.J. (1995). Structure of the catalytic region of human complement protease C1s: Study by chemical cross-linking and three-dimensional homology modeling. Biochemistry **34**, 7311-7321.

Rost B. & Sander C. (1996). Bridging the protein sequence-structure gap by structure predictions. Annu. Rev. Biophys. Biomol. Struct. **25**, 113-136.

Ruan K.-H., Milfeld K., Kulmacz R.J. & Wu K.K. (1994). Comparison of the construction of a 3-D model for human thromboxane synthase using P450cam and BM-3 as templates: Implications for the substrate binding pocket. Prot. Eng. **7**, 1345-1351.

Ruff-Jamison S. & Glenney J.R., Jr. (1993). Molecular modeling and site-directed mutagenesis of an anti-phosphotyrosine antibody predicts the combining site and allows the detection of higher affinity interactions. Prot. Eng. **6**, 661-668.

Ruffle S.V., Donnelly D., Blundell T.L. & Nugent J.H.A. (1992). A three-dimensional model of the photosystem II reaction centre of *Pisum sativum*. Photosynthesis Res. **34**, 287-300.

Rufino S.D., Donate L.E., Canard L.H.J. & Blundell T.L. (1997). Predicting the conformational class of short and medium size loops connecting regular secondary structures: Application to comparative modelling. J. Mol. Biol. **267**, 352-367.

Russell R.B. & Barton G.J. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts, secondary structure and accessibility. J. Mol. Biol. **244**, 332-350.

Saldanha J. & Mahadevan D. (1991). Molecular model-building of amylin and α-calcitonin gene-related polypeptide hormones using a combination of knowledge sources. Prot. Eng. **4**, 539-544.

Sali A. (1995). Modelling mutations and homologous proteins. Curr. Opin. Biotech. **6**, 437-451.

Sali A. & Blundell T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**, 779-815.

Sali A., Matsumoto R., McNeil H.P., Karplus M. & Stevens R.L. (1993). Three-dimensional models of four mouse mast cell chymases. J. Biol. Chem. **268**, 9023-9034.

Samsom M.S.P. & Kerr I.D. (1993). Influenza virus M_2 protein: A molecular modelling study of the ion channel. Prot. Eng. **6**, 65-74.

Samudrala R. & Moult J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. **275**, 895-916.

Saqi M.A.S., Bates P.A. & Sternberg M.J.E. (1992). Towards an automatic method of predicting protein structure by homology: An evaluation of suboptimal sequence alignments. Prot. Eng. **5**, 305-311.

Scarborough P.E., Guruprasad K., Topham C., Richo G.R., Conner G.E., Blundell T.L. & Dunn B.M. (1993). Exploration of subsite binding specificity of human cathepsin D through kinetics and rule-based molecular modeling. Protein Sci. **2**, 264-276.

Schiffer C.A., Caldwell J.W., Kollman P.A. & Stroud R.M. (1990). Prediction of homology protein structures based on conformational searches and energetics. Proteins **8**, 30-43.

Schoentgen F., Seddiqi N., Bucquoy S., Jollés P., Lemesle-Varloot L., Provost K. & Mornon J.-P. (1992). Main structural and functional features of the basic cytosolic bovine 21 kDa protein delineated through hydrophobic cluster analysis and molecular modelling. Prot. Eng. **5**, 295-303.

Schrauber H., Eisenhaber F. & Argos P. (1993). Rotamers: to be or not to be? An analysis of amino acid side chain conformations in globular proteins. J. Mol. Biol. **230**, 592-612.

Scully J.L. & Evans D.R. (1991). Comparative modeling of mammalian aspartate transcarbamylase. Proteins **9**, 191-206.

Secrist III J.A., Niwas S., Rose J.D., Babu Y.S., Bugg C.E., Erion M.D., Guida W.C., Ealick S.E. & Montgomery J.A. (1993). Structure-based design of inhibitors of purine nucleoside phosphorylase. 2. 9-alicyclic and 9-heteroalicyclic derivatives of 9-deazaguanine. J. Med. Chem. **36**, 1847-1854.

Selvaggini C., Salmona M. & de Gioia L. (1995). Manganese peroxidase from *Phanerochaete chrysosporium*. A homology-based molecular model. Eur. J. Biochem. **228**, 955-961.

Sham H.L., Bolis G., Stein H.H., Fesik S.W., Marcotte P.A., Plattner J.J., Rempel C.A. & Greer J. (1988). Renin inhibitors. Design and synthesis of a new class of conformationally restricted analogues of angiotensinogen. J. Med. Chem. **31**, 284-295.

Shenkin P.S., Farid H. & Fetrow J.S. (1996). Prediction and evaluation of side-chain conformations for protein backbone structures. Proteins **26**, 323-352.

Shenkin P.S., Yarmush D.L., Fine R.M., Wang H. & Levinthal C. (1987). Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers **26**, 2053-2085.

Shirai H., Kidera A. & Nakamura H. (1996). Structural classification of CDR-H3 in antibodies. FEBS Lett. **399**, 1-8.

Shoichet B.K., Stroud R.M., Santi D.V., Kuntz I.D. & Perry K.M. (1993). Structure-based discovery of inhibitors of thymidylate synthase. Science **259**, 1445-1450.

Sibanda B.L., Blundell T., Hobart P.M., Fogliano M., Bindra J.S., Dominy B.W. & Chirgwin J.M. (1984). Computer graphics modelling of human renin. FEBS Lett. **174**, 102-111.

Siezen R.J., de Vos W.M., Leunissen J.A.M. & Dijkstra B.W. (1991). Homology modelling and protein en-

gineering strategy of subtilases, the family of subtilisin-like serine proteinases. Prot. Eng. **4**, 719-737.

Siezen R.J., Creemers J.W.M. & van de Ven W.J.M. (1994). Homology modelling of the catalytic domain of human furin. A model for the eukaryotic subtilisin-like proprotein convertases. Eur. J. Biochem. **222**, 255-266.

Siezen R.J., Rollema H.S., Kuipers O.P. & de Vos W.M. (1995). Homology modelling of the *Lactococcus lactis* leader peptidase NisP and its interaction with the precursor of the lantibiotic nisin. Prot. Eng. **8**, 117-125.

Signor G., Vita C., Fontana A., Frigerio F., Bolognesi M., Toma S., Gianna R., de Gregoriis E. & Grandi G. (1990). Structural features of neutral protease from *Bacillus subtilis* deduced from model-building and limited proteolysis experiments. Eur. J. Biochem. **189**, 221-227.

Singh J., Dobrusin E.M., Fry D.W., Haske T., Whitty A. & McNamara D.J. (1997). Structure-based design of a potent, selective, and irreversible inhibitor of the catalytic domain of the erbB receptor subfamily of protein tyrosine kinases. J. Med. Chem. **40**, 1130-1135.

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.

Smith-Gill S.J., Mainhart C., Lavoie T.B., Feldmann R.J., Drohan W. & Brooks B.R. (1987). A three-dimensional model of an anti-lysozyme antibody. J. Mol. Biol. **194**, 713-724.

Sönnichsen F.D., Sykes B.D. & Davies P.L. (1995). Comparative modeling of the three-dimensional structure of type II antifreeze protein. Protein Sci. **4**, 460-471.

Southerland W.M. (1994). A molecular model of the folate binding site of *Pneumocystis carinii* dihydrofolate reductase. J. Comput.-Aided Mol. Design **8**, 113-122.

Sowdhamini R., Ramakrishnan C. & Balaram P. (1993). Modelling multiple disulphide loop containing polypeptides by random conformation generation. The test cases of α-conotoxin GI and endothelin I. Prot. Eng. **6**, 873-882.

Srinivasan N. & Blundell T.L. (1993). An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. Prot. Eng. **6**, 501-512.

Srinivasan R. & Rose G.D. (1995). LINUS: A hierarchic procedure to predict the fold of a protein. Proteins **22**, 81-99.

Srinivasan S., March C.J. & Sudarsanam S. (1993). An automated method for modeling proteins on known templates using distance geometry. Protein Sci. **2**, 277-289.

Steinkasserer A., Barlow P.N., Willis A.C., Kertesz Z., Campbell I.D., Sim R.B. & Norman D.G. (1992). Activity, disulphide mapping and structural modelling of the fifth domain of human b_2-glycoprotein I. FEBS Lett. **313**, 193-197.

Sticht H., Gallert K.-C., Krauss G. & Rösch P. (1997). Homology modeling of adenylosuccinate synthetase from Saccharomyces cerevisiae reveals a possible binding region for single-stranded ARS sequences. J. Biomol. Struct. Dynam. **14**, 667-675.

Strahs D. & Weinstein H. (1997). Comparative modeling of molecular dynamics studies of the δ, κ, and μ opioid receptors. Prot. Eng. **10**, 1019-1038.

Straßburger W., Wollmer A., Pitts J.E., Glover I.D., Tickle I.J., Blundell T.L., Steffens G.J., Günzler W.A., Ötting F. & Flohé L. (1983). Adaptation of plasminogen activator sequences to known protease structures. FEBS Lett. **157**, 219-223.

Sudarsanam S. (1998). Structural diversity of sequentially identical subsequences of proteins: Identical octapeptides can have different conformations. Proteins **30**, 228-231.

Sudarsanam S., DuBose R.F., March C.J. & Srinivasan S. (1995). Modeling protein loops using a ϕ_{i+1}, ψ_i dimer database. Protein Sci. **4**, 1412-1420.

Sudarsanam S., March C.J. & Srinivasan S. (1994). Homology modeling of divergent proteins. J. Mol. Biol. **241**, 143-149.

Summers N.L. & Karplus M. (1989). Construction of side-chain in homology modelling application to the C-terminal lobe of rhizopuspepsin. J. Mol. Biol. **210**, 785-811.

Sutcliffe M.J., Hayes F.R.F. & Blundell T.L. (1987). Knowledge based modelling of homologous proteins, part II: Rules for the conformations of substituted sidechains. Prot. Eng. **1**, 385-392.

Svensson B., Vass I., Cedergren E. & Styring S. (1990). Structure of donor side components in the photosystem II predicted by computer modelling. EMBO J. **9**, 2051-2059.

Swindells M.B. & Thornton J.M. (1991). Modelling by homology. Curr. Opin. Struct. Biol. **1**, 219-223.

Szilágyi A. & Závodszky P. (1995). Structural basis for the extreme thermostability of D-glyceraldehyde-3-phosphate dehydrogenase from *Thermotoga maritima*: Analysis based on homology modelling. Prot. Eng. **8**, 779-789.

Szklarz G.D. & Halpert J.R. (1997). Molecular modeling of cytochrome P450 3A4. J. Comput.-Aided Mol. Design **11**, 265-272.

Szklarz G.D., He Y.A. & Halpert J.R. (1995). Site-directed mutagenesis as a tool for molecular modeling of cytochrome P450 2B1. Biochemistry **34**, 14312-14322.

Tanimura R., Kidera A. & Nakamura H. (1994). Determinants of protein side-chain packing. Protein Sci. **3**, 2358-2365.

Teeter M.M., Froimowitz M., Stec B. & DuRand C.J. (1994). Homology modeling of the dopamine D_2 receptor and its testing by docking of agonists and tricyclic antagonists. J. Med. Chem. **37**, 2874-2888.

Terry C.J. & Blake C.C.F. (1992). Comparison of the modelled thyroxine binding site in TBG with he experimentally determined site in transthyretin. Prot. Eng. **5**, 505-510.

Thomä N.H. & Leadlay P.F. (1996). Homology modeling of human methylmalonyl-CoA mutase: A structural basis for point mutations causing methylmalonic aciduria. Protein Sci. **5**, 1922-1927.

Tilley J.W., Chen L., Fry D.C., Emerson S.D., Powers G.D., Biondi D., Varnell T., Trilles R., Guthrie R., Mennona F., Kaplan G., LeMahieu R.A., Carson M., Han R.-J., Liu C.-M., Palermo R. & Ju G. (1997). Identification of a small molecule inhibitor of the IL2/IL-2Ra receptor interaction which binds to IL-2. J. Amer. Chem. Soc. **119**, 7589-7590.

Toma K., Yamamoto S., Deyashiki Y. & Suzuki K. (1987). Three-dimensional structure of protein C inhibitor predicted from structure of α1-antitrypsin with computer graphics. Prot. Eng. **1**, 471-475.

Topham C.M., Overington J., Kowlessur D., Thomas M., Thomas E.W. & Brocklehurst K. (1990a). Investigation of mechanistic consequences of natural structural variation within the cysteine proteinases by knowledge-based modelling and kinetic methods. Biochem. Soc. Trans. **18**, 579-580.

Topham C.M., Overington J., O'Driscoll M., Salih E., Thomas M., Thomas E.W. & Brocklehurst K. (1990b). Three-dimensional structure of a B-type chymopapain. Biochem. Soc. Trans. **18**, 933-934.

Topham C.M., Salih E., Frazao C., Kowlessur D., Overington J., Thomas M., Brocklehurst S.M., Patel M., Thomas E.W. & Brocklehurst K. (1991). Structure-function relationships in the cysteine proteinases actinidin, papain and papaya proteinase W. Biochem J. **280**, 79-92.

Tramontano A. & Lesk A.M. (1992). Common features of the conformations of antigen-binding loops in immunoglobulins and application to modelling loop conformations. Proteins **13**, 231-245.

Travers P., Blundell T.L., Sternberg M.J.E. & Bodmer W.F. (1984). Structural and evolutionary analysis of HLA-D-region products. Nature **310**, 235-238.

Trumpp-Kallmeyer S., Hoflack J., Bruinvels A. & Hibert M. (1992). Modeling of G-protein-coupled receptors: Application to dopamine, adrenaline, serotonin, acetylcholine, and mammalian opsin receptors. J. Med. Chem. **35**, 3448-3462.

Tufféry P., Etchebest C., Hazout S. & Lavery R. (1991). A new approach to the rapid determination of protein side chain conformations. J. Biomol. Struct. Dynam. **8**, 1267-1289.

Tufféry P., Etchebest C. & Hazout S. (1997). Prediction of protein side chain comformations: A study on the influence of backbone accuracy on conformation stability in the rotamer space. Prot. Eng. **10**,

361-372.

Unger R., Harel D., Wherland S. & Sussman J.L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins **5**, 355-373.

van Gelder C.W.G., Leusen F.J.J., Leunissen J.A.M. & Noordik J.H. (1994). A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. Proteins **18**, 174-185.

van Gunsteren W.F. & Berendsen H. (1987). Groningen molecular simulation (GROMOS) library manual., Biomos. Nijenborgh 16, Groningen, The Netherlands.

Varney M.D., Marzoni G.P., Palmer C.L., Deal J.G., Webber S., Welsh K.M., Bacquet R.J., Barlett C.A., Morse C.A., Booth C.L.J., Herrmann S.M., Howland E.F., Ward R.W. & White J. (1992). Crystal-structure-based design and synthesis of benz[cd]indole-containing inhibitors of thymidylate synthase. J. Med. Chem. **35**, 663-676.

Vihinen M., Lundin M. & Baltscheffsky H. (1992). Computer modeling of two inorganic pyrophosphatases. Biochem. Biophys. Res. Commun. **186**, 122-128.

Vihinen M. (1994). Modeling of prostate specific antigen and human glandular kallikrein structures. Biochem. Biophys. Res. Commun. **204**, 1251-1256.

Villoutreix B.O., Getzoff E.D. & Griffin J.H. (1994). A structural model for the prostate disease maker, human prostate-specific antigen. Protein Sci. **3**, 2033-2044.

Villoutreix B.O. & Dahlback B. (1998). Structural investigation of the A domains of human blood coagulation factor V by molecular modeling. Protein Sci. **7**, 1317-1325.

Villoutreix B.O., Härdig Y., Wallqvist A., Covell D.G., de Frutos P.G. & Dahlbäck B. (1998). Structural investigation of C4b-binding protein by molecular modeling: Location of putative binding sites. Proteins **31**, 391-405.

Vinals C., de Bolle X., Depiereux E. & Feytmans E. (1995). Knowledge-based modeling of the D-lactate dehydrogenase three-dimensional structure. Proteins **21**, 307-318.

Viswanathan M., Anchin J.M., Droupadi P.R., Mandal C., Linthicum D.S. & Subramaniam S. (1995). Structural predictions of the binding site architecture for monoclonal antibody NC6.8 using computer-aided molecular modeling, ligand binding, and spectroscopy. Biophys. J. **69**, 741-753.

von Itzstein M., Wu W.-Y., Kok G.B., Pegg M.S., Dyason J.C., Jin B., Phan T.V., Smythe M.L., White H.F., Oliver S.W., Colman P.M., Varghese J.N., Ryan D.M., Woods J.M., Bethell R.C., Hotham V.J., Cameron J.M. & Penn C.R. (1993). Rational design of potent sialidase-based inhibitors of influenza virus replication. Nature **363**, 418-423.

Voorhorst W.G.B., Warner A., de Vos W.M. & Siezen R.J. (1997). Homology modelling of two subtilisin-like serine proteases from the hyperthermophilic archaea *Pyrococcus furiosus* and *Thermococcus stetteri*. Prot. Eng. **10**, 905-914.

Wampler J.E., Bradley E.A., Stewart D.E. & Adams M.W.W. (1993). Modeling the structure of *Pyrococcus furiosus* rubredoxin by homology to other X-ray structures. Protein Sci. **2**, 640-649.

Wang L.-H., Matijevic-Aleksic N., Hsu P.-Y., Ruan K.-H., Wu K.K. & Kulmacz R.J. (1996). Identification of thromboxane A_2 synthase active site residues by molecular modeling-guided site-directed mutagenesis. J. Biol. Chem. **271**, 19970-19975.

Warme P.K., Momany F.A., Rumball S.V., Tuttle R.W. & Scheraga H.A. (1974). Computation of structures of homologous proteins. α-lactalbumin from lysozyme. Biochemistry **13**, 768-782.

Weber I.T., Miller M., Jaskólski M., Leis J., Skalka A.M. & Wlodawer A. (1989). Molecular modeling of the HIV-1 protease and its substrate binding site. Science **243**, 928-931.

Weiner S.J., Kollman P.A., Nguyen D.T. & Case D.A. (1986). An all atom force field for simulations of proteins and nucleic acids. J. Comp. Chem. **7**, 230-252.

West S., Bamborough P. & Tully R. (1993). Tertiary structure of calcineurin B by homology modeling. J. Mol. Graph. **11**, 47-52.

Wibley J.E.A., McKie J.H., Embrey K., Marks D.S., Douglas K.T., Moore M.H. & Moody P.C.E. (1995). A homology model of the three-dimensional structure of human O^6-alkylguanine-DNA alkyltransferase based on the crystal structure of the C-terminal domain of the Ada protein from *Escherichia coli*. Anti-Cancer Drug Des. **10**, 75-95.

Wilson I.A., Haft D.H., Getzoff E.D., Tainer J.A., Lerner R.A. & Brenner S. (1985). Identical short peptide sequences in unrelated proteins can have different conformations: A testing ground for theories of immune recognition. Proc. Natl. Acad. Sci. USA **82**, 5255-5259.

Wistow G., Slingsby C. & Blundell T. (1981). Eye-lens proteins: The three-dimensional structure of β-crystallin predicted from monomeric γ-crystallin. FEBS Lett. **133**, 9-16.

Wodak S.J. & Rooman M.J. (1993). Generating and testing protein folds. Curr. Opin. Struct. Biol. **3**, 247-259.

Xiong J., Subramaniam S. & Govindjee. (1996). Modeling of the D1/D2 proteins and cofactors of the photosystem II reaction center: Implications for herbicide and bicarbonate binding. Protein Sci. **5**, 2054-2073.

Zemla A., Venclovas C., Reinhardt A., Fidelis K. & Hubbard T.J. (1997). Numerical criteria for the evaluation of ab initio prediction of protein structure. Proteins **Suppl. 1**, 140-150.

Zhang C. & DeLisi C. (1998). Estimating the number of protein folds. J. Mol. Biol. **284**, 1301-1305.

Zheng Q., Rosenfeld R., Vajda S. & Delisi C. (1993a). Determining protein loop conformation using scaling-relaxation techniques. Protein Sci. **2**, 1242-1248.

Zheng Q., Rosenfeld R., Vajda S. & Delisi C. (1993b). Loop closure via bond scaling and relaxation. J. Comp. Chem. **14**, 556-565.

Zvelebil M.J.J.M., Wolf C.R. & Sternberg M.J.E. (1991). A predicted three-dimensional structure of human cytochrome P450: Implications for substrate specificity. Prot. Eng. **4**, 271-282.

RESUME

La modélisation moléculaire est un ensemble de techniques qui permettent d'étudier la fonction d'une molécule grâce à la connaissance de sa structure tridimensionnelle. Ce mémoire présente en détail une de ces techniques : la modélisation par homologie.
La modélisation par homologie permet d'obtenir la structure tridimensionnelle d'une protéine (cible) lorsque la séquence de celle-ci est suffisamment proche de celle d'une protéine dont la structure tridimensionnelle (parente) a été déterminée expérimentalement (RX ou RMN). L'essence de la modélisation par homologie se trouve dans le fait que la séquence d'une protéine évolue plus vite que sa structure tridimensionnelle. Par conséquent, des protéines ayant une faible homologie de séquence peuvent néanmoins avoir une forte homologie structurale.
Cette technique consiste en six étapes : 1) reconnaissance d'un repliement adéquat (parent), 2) alignement des séquences entre la molécule cible et la molécule parente, 3) construction de la chaîne principale, 4) construction des chaînes latérales, 5) construction des insertions/suppressions et 6) optimisation du modèle. La recherche de repliements adéquats se fait par criblage des protéines présentes dans la PDB. Une technique récente consiste à identifier structuralement l'adéquation entre un repliement et une séquence particulière ("threading"). Une fois le repliement défini, l'étape la plus importante consiste à aligner les séquences de la protéine cible à la molécule identifiée. Cette étape est cruciale puisqu'elle définit les régions conservées entre les deux séquences ainsi que les limites des insertions et des suppressions. En pratique la construction de la chaîne principale et des chaînes latérales se font par simple transposition des coordonées cartésiennes des régions conservées. Pour les régions les moins conservées, souvent les boucles en surface, il est possible de recourir à des recherches de fragments peptidiques dans la base de données structurales ou d'utiliser des techniques de constructions *ab-initio* suivi de procédures stochastiques d'optimisation.
Quatre applications pratiques de la modélisation par homologie sont présentées. La première concerne la modélisation à moyenne résolution dans le but d'étudier le mécanisme de fonctionnement du cofacteur Va dans la cascade de coagulation sanguine. Le modèle présenté décrit l'orientation relative du cofacteur Va vis-à-vis de la membrane lipidique et permet d'émettre des hypothèses sur le mode de régulation de ce cofacteur grâce au positionnement de la protéine C activée. La seconde application concerne la modélisation à haute résolution dans le but d'étudier l'assemblage entre ligands et récepteurs pour deux systèmes : l'insertion du peptide consensus reconnu par l'antigène spécifique de la prostate et l'assemblage d'hydrocarbures aromatiques polycycliques dans le site de reconnaissance d'un anticorps. La troisième application concerne l'ingéniérie des protéines grâce au design d'immunotoxines de *Bacillus thuringiensis* en couplant plusieurs domaines de cette toxine à un fragment variable d'anticorps simple chaîne. Finalement, la quatrième application est de type bioinformatique où la modélisation moléculaire a permis d'établir un lien entre une famille de séquence (PAS) et un prototype structural (PYP). Un des grands défis à venir est l'attribution de fonction aux séquences issues des projets génomiques. L'utilisation de la modélisation est une technique raffinée pour identifier les homologies structurales entre des séquences divergentes et ainsi postuler leurs fonctions.
Bien que la modélisation par homologie soit une technique récente, elle a remporté de nombreux succès. Cependant, des problèmes persistent. Il est par exemple nécessaire de comprendre les changements conformationnels dans les protéines dans le but de pouvoir les anticiper lors de l'assemblage d'un ligand à son récepteur. Il est également important que la modélisation par homologie affine sa technique d'alignement de séquences et que les techniques stochastiques d'échantillonnage soient perfectionnées.
Peut-être que la modélisation moléculaire n'atteint pas encore l'état dans lequel on voudrait qu'elle soit ! Mais il est clair qu'il faut compter avec elle en Biologie et en Bioinformatique.