Title

The structural basis for highly specific protein-ligand complexes. I. Apolipoprotein-E and the LDL receptor. II. Engineering the substrate preferences of alpha-lytic protease

Author

Wilson, Charles,

Publication Date

1991

Peer reviewed|Thesis/dissertation

The structural basis for highly specific protein-ligand complexes. I. Apolipoprotein-E and the LDL receptor. II. Engineering the substrate preferences of alpha-lytic protease.

by

Charles Wilson

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

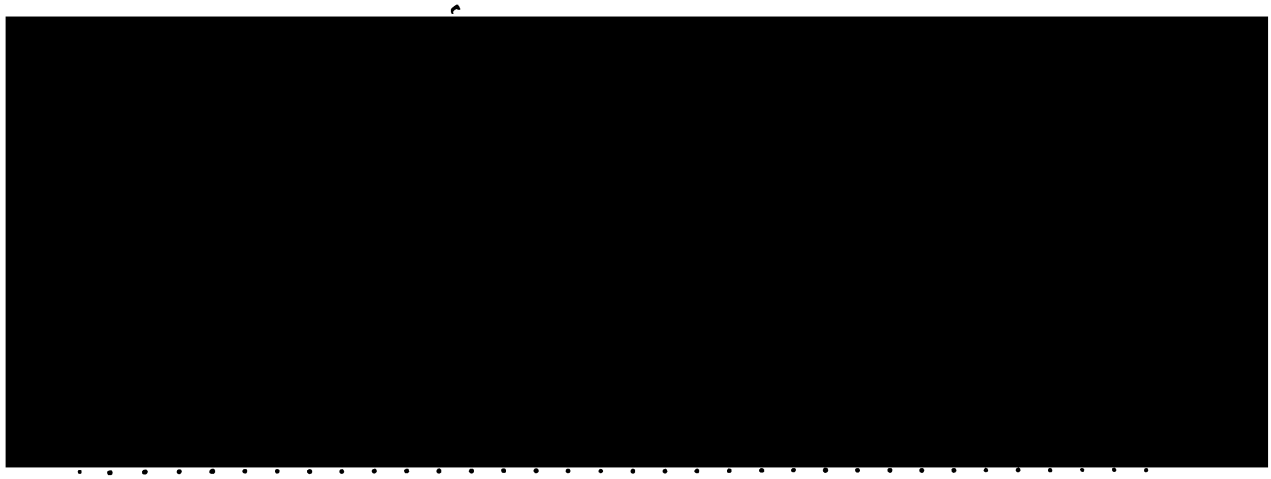
in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco



Date . . . . . University Librarian

Degree Conferred: 6/16/91

Dedicated to my parents, Keith and Diana,
and to my sweetheart, Lydia.

Those, O, so long four helices,
with leucine-zipper qualities
and $P2_12_12_1$ symmetry,
were, for years, the bane of me.

But thanks to dimethyl mercury,
and the awesome power of diffractometry,
I'm going to get my Ph.D.,
and thus ends the ode to apo-E.

–CW 1990

Acknowledgements

I would like to acknowledge the people at UCSF who have helped to make my graduate school experience a rich and memorable one. Most importantly, I would like to thank my mentor, David Agard. David has been a wonderful thesis advisor, encouraging me to pursue a wide variety of different projects while simultaneously providing the focus needed to finish some of them. I look forward to collaborating with David in the future to follow up on some of the ideas our work has suggested.

My progress on modelling the substrate specificity of α-lytic protease relied heavily upon the experimental work of others in the Agard laboratory, most notably Roger Bone, Joy Silen, and Jim Mace. Roger was an excellent teacher in the time that we overlapped at UCSF, getting me started on the apo-E crystallography and the substrate specificity calculations. Joy's work to engineer the expression plasmids for mutant enzymes and Jim's work to kinetically characterize these enzymes helped to provide the large body of data that was absolutely required for carrying out the empirical energy calculations.

Julie Ransom is the nicest administrator ever made, and her efforts to slice through the UCSF bureaucracy to get my pay checks were much appreciated. The Hertz Foundation has been very generous in its financial support of me throughout my graduate school career and, in particular, I would like to thank Dr. Talley for negotiating the transfer of my fellowship from Stanford to UCSF.

Brian Shoichet, Jason Swedlow, Inke Nathke, and Dave Yee helped to fill my time outside the lab with some of the most enjoyable times of my life, making it very hard to leave San Francisco with a modicum of sanity. The occasional trips to Tassahara, Vesuvio's, Steep Ravine, Baker Beach, and the Sierra foothills were welcome distractions from graduate school that kept my spirits up and made UCSF a great place to study.

Abstract

The structural basis for highly specific protein-ligand complexes: I. Apolipoprotein-E and the LDL receptor. II. Engineering the substrate preferences of alpha-lytic protease.

by

Charles Wilson

Many biological events are predicated upon the highly specific interaction between a protein and its appropriate ligand. Understanding the energetic factors that drive the formation of protein complexes is a key step towards the design of novel enzymes or drugs. This thesis addresses the structural basis for the high specificity of protein-ligand complexes, focussing on two independent model systems: 1) the interaction of human apolipoprotein-E (apo-E) with the low density lipoprotein (LDL) receptor, and 2) the specificity of the enzyme α-lytic protease for particular substrates. X-ray crystallography has shown the receptor binding domain of apo-E to be an unusually elongated four helix bundle with several basic residues involved in receptor recognition clustered on a single helix. This structure, the first determined for a mammalian apolipoprotein, provides an understanding of the important structure-function relationships for this and related proteins, and serves as a starting point for rational anti-atherosclerosis drug design. In the process of solving the structure of apo-E, a method for crystallographic phase refinement based

upon iterative electron density skeletonization was developed and shown to be useful when using incomplete atomic structures to solve the phase problem. A Monte Carlo dynamics procedure was developed to simulate protein folding and used to predict the folded structure of the apolipoprotein-binding domain of the LDL receptor. Work on α-lytic protease has resulted in the development of a computational method that accurately predicts the effects of mutagenesis on enzyme substrate specificity. The method employs a library of side chain rotamers to sample protein conformation space and a free energy force field that includes solvation effects to calculate relative binding energies. Using this algorithm, a computer-designed enzyme with altered substrate specificity has been synthesized. In the future, a newly-developed genetic screen for protease activity (based upon pro-inducer activation) may be employed to test the predictions of this algorithm. The same computational method has been applied to the problem of predicting side chain conformation for homologous proteins with good success.

Table of Contents

List of Tables

List of Figures

Preface

Most biological events are based upon highly specific protein-ligand interactions. To function properly, enzymes, receptors, gene regulators, and antibodies must recognize and bind with high affinity to their appropriate target (*i.e.* the proper substrate, hormone, operator sequence, or antigen). In many respects, a specific protein-ligand complex can be thought of in terms of a lock-and-key analogy. While most residues in a protein are not directly involved in ligand binding, they provide a relatively rigid framework that arranges a handful of amino acids in a precise orientation with respect to one another to form a ligand binding site. This positioning of key residues provides for the energetically-favorable hydrogen bonds, salt bridges, or van der Waals interactions that drive complex formation with the ligand. Understanding the basis for high specificity within a protein-ligand complex clearly requires an intimate knowledge of its three-dimensional structure. By characterizing and manipulating a number of complexes, it should be possible to quantitatively measure the contribution of the different interatomic forces that stabilize ligand binding. Ultimately, with this knowledge one should be able to engineer proteins that bind new ligands or design new drugs that interact specifically with a desired target protein.

This thesis addresses the structural basis for the high specificity of protein-ligand complexes, focussing on two independent model systems: 1) the interaction of human apolipoprotein-E with the low density lipoprotein (LDL) receptor, and 2) the specificity of the enzyme α-lytic protease for particular substrates. Work on apolipoprotein-E has led to the first crystal structure of a mammalian apolipoprotein (the class of proteins responsible for lipid and cholesterol transport), which will hopefully provide an understanding of the important structure-function relationships for this and related proteins, and serve as a starting point for rational anti-atherosclerosis drug design. Work on α-lytic protease has

resulted in the development of a computational method that accurately predicts the effects of mutagenesis on enzyme substrate specificity. Using this algorithm, the first computer-designed enzyme with altered specificity has been synthesized. In the future, the same algorithm will hopefully be applied to the engineering of many other enzymes. The following paragraphs outline and summarize the major points made in each of the subsequent chapters.

Chapters 1-6 describe structural studies of human apolipoprotein-E and its interaction with the LDL receptor. While there are several well characterized protein complexes, the clear physiological importance of this protein-receptor interaction makes it an obvious target for structural analysis. Apolipoprotein-E (apo-E) is a small protein found in the blood plasma that directs the receptor-mediated endocytosis of a number of lipoproteins, including high density lipoprotein (HDL), very low density lipoprotein (VLDL), and chylomicron remnants (reviewed by Mahley, 1988). In the absence of proper apo-E binding to cell surface receptors (such as the LDL receptor), lipoproteins accumulate in the bloodstream in a process that can eventually lead to the formation of atherosclerotic plaques. Characterization of the interaction between apo-E and the LDL receptor will thus help further elucidate the mechanism of cholesterol transport and is likely to aid the development of therapeutic agents for combating atherosclerosis. To this end, a major part of this thesis describes the crystallization and structure determination by x-ray diffraction of the LDL receptor-binding domain of human apo-E. Preliminary work towards predicting the structure of the ligand binding domain of the LDL receptor and the receptor-apo-E complex is also described.

Chapter one summarizes the biology of lipoproteins as it relates to apo-E and serves as an expanded introduction for chapter two, a discussion of the crystal structure of the LDL receptor binding domain of the human protein. This work has shown that the LDL receptor binding domain is an unusually long four helix bundle, made up by short (22-

amino acid) repeats corresponding to amphipathic alpha-helices. A combination of leucine zipper-type interactions in the hydrophobic core and numerous salt bridges at the protein surface appear to stabilize the folded conformation. Basic amino acids known to be important for LDL receptor binding are clustered in a single α-helix, arranged such that arginine and lysine side chains are largely solvent accessible, free for interaction with the receptor. The crystal structure suggests that electrostatic interactions (salt bridges) are likely to play a key role in generating apo-E's high affinity for the LDL receptor.

Chapter three describes the crystal structures of the two common mutant forms of human apo-E and relates their altered biological functioning to structural changes induced by single site amino acid substitutions. Apo-E is one of the most polymorphic human genes that has been characterized and serves as one of the best genetic markers of increased risk for cardiovascular disease (Davignon, 1988). One frequently-occuring mutation (the apo-E2 mutant) results in a protein with dramatically reduced LDL receptor binding. Crystallographic analysis suggests that this mutant induces a large conformational change which, by disrupting both electrostatic and steric interactions with the receptor, may be responsible for reduced binding. A second common mutation (apo-E4) leaves receptor binding intact but significantly alters the lipoprotein binding properties of apo-E, increasing the relative affinity for VLDL. This effect is surprising since lipoprotein binding is largely mediated by the C-terminal domain, yet the E4 mutation corresponds to a single amino acid substitution in the N-terminal domain. The crystal structure of this mutation shows minimal conformational changes, implying an indirect mechanism for this mutant's dysfunction. One possibility is that the mutation changes the interaction between the N- and C-terminal domains such that the lipid binding properties of the C-terminal domain are altered. If this is the case, the crystal structure suggests that inter-domain communication within a protein can be accomplished in the absence of significant structural changes as have been observed for allosteric enzymes.

X-ray crystallography remains the best experimental technique for obtaining the high resolution structural information needed to understand the functioning of macromolecules. While the rate of x-ray structure determination seems to be constantly increasing, a crystallography project can very often be thwarted by the lack of heavy atom derivatives, usually required for finding a solution to the crystallographic phase problem. Indeed, most of the time and effort spent in solving the structure of apo-E involved the random screening of approximately 50 heavy metal compounds, searching for a suitable isomorphous derivative. Any computational method capable of solving the phase problem in the absence of experimental phase information would thus greatly simplify and accelerate the process of solving protein crystal structures. After collecting a native data set for apo-E, it was noticed that an idealized four-helix bundle reproduced some of the important features of the native diffraction pattern. Chapter four describes efforts to develop an algorithm capable of using the phase information in such a simplified, partial model as a starting point for refinement to the true phases. The method developed is based on iterative skeletonization of electron density maps, and as with solvent flattening techniques, repeatedly applies constraints to the phases in both real space and reciprocal space representations of the scattering. Simple test cases using data calculated from helical proteins suggests that it has a much larger radius of convergence than conventional solvent flattening and may thus become a generally useful scheme for structure refinement.

In the absence of a crystal structure of the LDL receptor complexed to apo-E, a predicted structure of the receptor domain implicated in apolipoprotein binding has been constructed. Chapter five describes a Monte Carlo dynamics procedure developed for the simulation of protein folding and for predicting protein tertiary structure solely from primary sequence information. Simulated annealing, using empirical potentials derived from known protein structures to evaluate alternate conformations, has been applied to the small disulfide-rich proteins crambin and BPTI (bovine pancreatic trypsin inhibitor) with a

fair degree of success. Chapter six describes the application of this algorithm to a 46-amino acid sequence derived from the LDL receptor. Deletion of this peptide dramatically reduces apo-E binding to the receptor, suggesting that this domain may directly mediate apolipoprotein binding (Russell, 1989). A series of computer experiments using the folding algorithm have suggested a structure for the LDL receptor repeat that is consistent with the limited experimental information obtained from the synthetic refolded peptide. With the predicted structure of the receptor's binding domain and the crystal structure of the receptor's ligand, one can start to model the interaction between the two species and begin to understand the structural basis for high affinity binding.

In the second half of this thesis, chapters 7-9 describe the development of computational and experimental methods for designing enzymes with altered substrate specificity. Most enzymes are designed to efficiently utilize only a small number of compounds as substrates for catalysis. A large body of recent experimental work has demonstrated that enzyme substrate specificity can be altered dramatically by protein engineering without destroying catalytic efficiency (reviewed by Wilson, 1991a). Much effort has been spent on the development of methods that can predict the effects of site directed mutagenesis on enzyme substrate specificity and thereby allow the rational design of enzymes with desired altered properties. Pioneering work by Kollman, Warshel, and others has shown that the effects of protein engineering can be accurately reproduced by the computational technique of free energy perturbation (Bash, 1987; Warshel, 1988). While quite impressive in terms of its accuracy, free energy perturbation is an extraordinarily computer-intensive method and therefore does not lend itself well to enzyme design (given that thousands of different mutant enzymes may need to be evaluated before one with the desired activity is discovered). Very rapid computational alternatives to free energy perturbation seem essential to solve the general problem of enzyme design.

For several reasons, the bacterial enzyme α-lytic protease serves as an ideal model system for studying and engineering substrate specificity. Contacts between the substrate and the enzyme in the binding pocket are mediated entirely by side chain atoms rather than by main chain atoms. As a consequence, site-direct mutagenesis of the binding pocket residues can be used to dramatically alter the enzyme-substrate interface. A large body of data for the substrate binding preferences of several binding pocket mutants has been gathered in the Agard laboratory (Bone, 1989). Kinetic constants for substrate catalysis have been complemented by extensive structural characterization of the mutant-substrate complexes. In particular, crystallographic studies of the binding of peptide boronic acids in the protease active site provide some of the best available models of enzyme-substrate transition state complexes (Bone, 1987). This data base of experimental observations provides a rigorous test for any algorithm designed to predict the effects of mutagenesis on substrate specificity.

Chapter seven describes the development, parameterization, and testing of a computational method for enzyme design. Using a library of side chain rotamers to sample conformation space and an empirical free energy force field that takes into account solvation effects, the relative effect of a substrate binding pocket mutation on the catalytic transition state binding energy can be rapidly and accurately estimated. The method yields a surprisingly good agreement between calculated and experimentally-determined relative binding energies for over forty different enzyme-substrate combinations. Using the algorithm, an α-lytic protease mutant predicted to be highly active and highly specific for a non-natural substrate has been designed. Experimental characterization of the mutant supports the predictions and suggests that the algorithm may be a generally useful tool for protein engineering.

As a complement to the computational method described above, chapter eight describes the development of a genetic screen/selection for experimentally isolating

enzymes with dramatically altered substrate specificity. The screen is based on the concept of pro-inducer conversion — by providing a cell culture with an inactive 'pro-inducer' for a reporter gene, individual cells that express an enzyme capable of converting the pro-inducer to an active form can be detected. The *dsdA* operon (induced by D-serine) has been co-opted to generate a protease-specific screen. This approach is ideally suited for substrate specificity studies since proteases are selected on the basis of their ability to cleave a well-defined substrate (a D-serine-containing hexapeptide) at a single site. The screen functions well with liquid cultures and by optimizing experimental conditions, it should perform with cells grown on solid media. By applying the screen to evaluate the results of combinatorial mutagenesis, one may be able to isolate enzymes with specificity for substrates that are not cleaved by any of the designed mutant proteases (*e.g.* identifying lysine-specific enzymes). Using a combination of computer-based methods and genetic screening, it should be possible to design and isolate highly specific enzymes for special purposes.

Chapter nine describes the application of the same computational method developed for enzyme design to the problem of predicting side chain conformation in the course of homology-modelling protein structure. Using a library of rotamers to represent the conformations available to side chains, this problem is reduced to one of identifying the correct rotamer at each position within the protein. To accomplish this, an algorithm has been developed which evaluates all combinations of rotamers within local clusters of amino acids, and adds the side chain combination with the lowest overall energy to the protein backbone. Tests with several pairs of homologous proteins are encouraging, suggesting that in a true modelling exercise, one should be able to predict the position of side chain atoms to within 1.5 Å of their true positions.

Each chapter contains its own introduction, methods, results, and discussion sections. Chapters two, five, and seven have been published or are in press (Wilson,

1989; Wilson, 1991b; Wilson, 1991c). Chapters three, four, and nine have been submitted for publication.

References

Bone, R., Shenvi, A.B., Kettner, C.A., & Agard, D.A. (1987). Serine protease mechanism: structure of an inhibitory complex of alpha-lytic protease and a tightly bound peptide boronic acid. *Biochemistry*, **27**, 7609-7614.

Bone, R., Silen, J.L., & Agard, D.A. (1989). Structural plasticity broadens the specificity of an engineered protease. *Nature*, **339**, 191-195.

Davignon, J., Gregg, R.E., & Sing, C.F. (1988). Apolipoprotein E polymorphism and atherosclerosis. *Arteriosclerosis*, **8**, 1-21.

Mahley, R.W. (1988). Apolipoprotein-E: cholesterol transport protein with expanding role in cell biology. *Science*, **240**, 622-630.

Russell, D.W., Brown, M.S., & Goldstein, J.L. (1989). Different combinations of cysteine-rich repeats mediate binding of low density lipoprotein receptor to two different proteins. *J. Biol. Chem.* **264**, 21682-21688.

Warshel, A., Sussman, F., & Hwang, J.K. (1988). Evaluation of catalytic free energies in genetically modified proteins. *J. Mol. Biol.* **201**, 139-159.

Wilson, C., & Doniach, S. (1989). A computer model to dynamically simulate protein folding: studies with crambin. *Proteins: structure, function, and genetics*, **6**, 193-209.

Wilson, C. & Agard, D.A. (1991). Engineering protein substrate specificity. *Current opinion in structural biology*. in the press.

Wilson, C., Mace, J.E., & Agard, D.A. (1991). A computational method for the design of enzymes with altered substrate specificity. *J. Mol. Biol.* in the press.

Wilson, C., Wardell, M.R., Weisgraber, K.H., Mahley, R.W., & Agard, D.A. (1991). The three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein-E. *Science*. in the press.

Chapter 1 :

The biology of lipoproteins and apolipoprotein-E

Introduction

Lipids (including triglycerides, fatty acids, phospholipids, and cholesterol) serve several important biochemical functions and are therefore exogenously required by most types of cells. Because of their generally poor aqueous solubility, lipids must be transported to tissues in the form of specialized particles called lipoproteins. All lipoproteins share a common structure in which a core of non-polar lipids is covered by a surface monolayer of phospholipids, embedded with lipoprotein-specific proteins (apolipoproteins). Lipoproteins have been extensively studied in the last thirty years because of their impact on the progression of atherosclerosis, the major cause of death in the United States. This chapter summarizes basic lipoprotein biology with particular emphasis on the role of apolipoprotein-E to their metabolism.

Lipoprotein classes

Lipoproteins are quite heterogeneous and are generally classified into five major categories on the basis of particle density (chylomicrons, VLDL, IDL, LDL, and HDL). Because of the radically different densities of lipid ($d < 1.0$) and protein ($d \approx 1.3$), particle density is a sensitive function of the overall lipid : protein ratio. Whereas the size of the protein component is relatively constant between lipoproteins, the volume of the lipid core varies widely and thus determines the overall particle density. The density classification accurately separates lipoproteins on the basis of their tissue of origin and by the types of apolipoproteins that they contain on their surface. The key properties of the five major classes are described below, tracing in order the flow of lipids through the body (summarized in figure 1.1).

Chylomicrons, produced by the intestine from dietary fat, are the lowest density lipoproteins ($d<1.006$) and have the largest lipid cores. The particles are very large

(diameter > 1000 Å) and therefore scatter light, giving the blood a milky appearance following a lipid-rich meal. Chylomicrons contain several different apolipoproteins, including apo-B48, apo-AI, apo-AIV, apo-CI, apo-CII, apo-CIII, and apo-E. The particles are broken down by lipoprotein lipase to yield chylomicron remnants, a class of small lipoproteins which are efficiently cleared from the blood by the liver. In the process of conversion to remnants, the phospholipid monolayer on the surface of chylomicrons is sluffed off, yielding nascent phospholipid disks which are believed to serve as the precursors for a fraction of high density lipoproteins (HDL) (Ruys, 1989). Apo-E may mediate the clearance of remnants from the plasma by binding the LDL receptor-related protein (LRP) on hepatocytes (Beisiegel, 1989). Triglycerides obtained from chylomicron remnants, together with cholesterol and additional apolipoproteins, are repackaged by the liver and secreted into the bloodstream in the form of VLDL (very low density lipoproteins), the second lowest density class of particles.

Like chylomicrons, VLDL are characterized by a large triglyceride-rich core (diameter ≈800 Å) and by a surface containing many different types of apolipoproteins. Whereas chylomicrons contain apolipoprotein-B48, VLDL carry the alternatively-spliced full length apolipoprotein-B100, which unlike its shorter derivative is capable of recognizing the LDL receptor with high affinity. In addition to apo-B, VLDL carry ≈ 20 molecules each of apo-E which direct their cellular uptake (the single molecule of apo-B does not seem to be in an active receptor-binding conformation in VLDL (Knott, 1986)). Apo-CII on the VLDL surface serves as an activating cofactor for lipoprotein lipase, enhancing the release of fatty acids from the lipid core. VLDL catabolism results transiently in the formation of VLDL remnants and intermediate density lipoproteins (IDL). Approximately 50% of secreted VLDL are cleared from the plasma by the liver following lipolysis. The remainder are processed more extensively, continuing to shrink in size, losing their triglycerides and shedding their apolipoproteins. Several hours after secretion

into the plasma, VLDL are fully converted into low density lipoproteins (LDL, 1.019 < d < 1.063), the major cholesterol-carrying lipoprotein fraction in the blood.

The core of an LDL particle (≈200 Å wide) is almost exclusively cholesterol and cholesterol ester while its protein coat contains a single molecule of apo-B100. In contrast to VLDL, LDL circulates in the bloodstream for several days. Most LDL is cleared from the plasma by apo-B-mediated binding to the LDL receptor (conversion from VLDL to LDL apparently induces a conformational change in apo-B which activates receptor binding). Oxidized or in other ways modified LDL can also be recognized and internalized via the macrophage scavenger receptor (Brown, 1983). Because this receptor is unregulated, macrophages are able to overload themselves with cholesterol, ultimately resulting in their morphogenesis into foam cells. Foam cells catalyze the maturation of atheromas into atherosclerotic lesions, and the role of the scavenger receptor in the development of atherosclerosis is thus being studied intensively (Kodama, 1990).

High density lipoproteins (HDL) are the densest, smallest lipoproteins (1.063 < d < 1.21, radius ≈70-100 Å). Most HDL are synthesized by hepatic cells although some fraction is derived from extrahepatic cells and from degraded chylomicrons. Apolipoprotein-AI forms ≈70% of the protein component of HDL and its major function involves stabilization of the particle structure. While apo-E is only a minor component of HDL, a large fraction (≈50%) of circulating apo-E is associated with this class of lipoprotein. The subclass of HDL which contains apo-E (HDL_1) appears to be involved in reverse cholesterol transport, a process whereby extrahepatic cells target their excess cholesterol to the liver for excretion (reviewed by Mahley, 1984). Apo-E apparently mediates the binding of HDL to the liver, perhaps via the LDL receptor-related protein. In most animals, this pathway is probably the major mechanism for removing excess cholesterol from tissues. In primates, however, this process is accomplished by the action of cholesterol ester transport protein (CETP), which catalyzes the transformation:

$$(HDL+CE) + (VLDL+TG) \rightarrow (HDL+TG) + (VLDL+CE)$$

(CE: cholesterol ester, TG: triglyceride). As a result, cholesterol-laden HDL are able to return their cholesterol to the LDL-receptor-regulated pool of lipoproteins. Interestingly, the apo-E-mediated pathway for reverse cholesterol transport seems to be more efficient and effective than the CETP-dependent pathway that normally operates in man. CETP-deficiency has been observed in a Japanese subpopulation; people lacking this protein show elevated HDL (enriched for apo-E), decreased LDL, and very few cases of atherosclerosis (Hoshino, 1989).

Apolipoprotein structure

Apo-E is one of a large family of evolutionarily-related apolipoproteins that also includes apo-AI, apo-AII, apo-AIV, apo-CI, apo-CII, and apo-CIII (Luo, 1986). The genes for these proteins are distributed in three clusters in the human genome: apo-AI, apo-CIII and apo-AIV are located on chromosome 11, apo-AII is on chromosome 1, while apo-CI, apo-CII, and apo-E are found on chromosome 19. On the basis of nucleotide conservation, these genes are postulated to have arisen from a single precursor apolipoprotein ≈600 million years ago, at approximately the same time that vertebrates diverged from invertebrates (Luo, 1986). These proteins share a similiar intron-exon structure, with the mature domains of most proteins consisting of one small N-terminal exon and a larger C-terminal exon. All of the apolipoproteins in this family contain 22-amino acid-long internal sequence repeats. The number of repeats varies considerably between the different proteins; while apo-CI contains only the amino-terminal half of one repeat, apo-AIV contains 13 repeats. The degree of conservation between repeats within a protein also differs significantly; the repeats in apo-AIV are virtually identical to one another whereas those in apo-E are barely detectable. Since they were first observed, the

apolipoprotein repeats have been predicted to form amphipathic helices (McLachlan, 1977). Figure 1.2 shows the first repeat from apo-E projected onto a helical wheel. One face of the helix is highly charged while the other contains only hydrophobic amino acids. Circular dichroism has shown that most apolipoproteins do have a significant amount of alpha-helical structure, supporting the amphipathic helix model for the repeats.

Most apolipoproteins are only sparingly soluble in aqueous solution and are found tightly associated with lipoprotein particles in the plasma. As a consequence, structural characterization of the apolipoproteins has been quite difficult. Apolipoprotein-E is somewhat unique in that it exists in solution as a stable tetramer — this property has made it a useful model system for structural studies. A series of biophysical experiments by John Wetterau, Lon Aggerbeck, and others has established that apo-E is a two-domain protein, with a large N-terminal domain and a smaller C-terminal domain separated by a short, flexible linker (Wetterau, 1988; Aggerbeck, 1988). Proteolysis of apo-E by thrombin releases 22-kDa and 10-kDa fragments corresponding to these two domains. The N-terminal fragment, responsible for the LDL receptor-binding properties of apo-E, behaves as a typical globular protein. Its free energy of stabilization as determined by guanidinium denaturation is ≈10 Kcal/mole, a typical value for a protein of its size. This domain is soluble to high concentrations (>10 mg/ml) and shows no tendency to oligomerize or aggregate in solution. These properties contrast strongly with those of the C-terminal fragment, which behaves much more like other apolipoproteins. This domain, which is responsible for apo-E binding to lipoprotein particles, has low stability in the absence of lipid, $\Delta G_{unfold} \approx 4$ Kcal/mole, and hydrodynamic studies suggest that it spontaneously forms oligomers in solution and may thus responsible for the tetramerization of the intact soluble form of apo-E.

The effect of lipid binding on apo-E structure has been examined by several different methods. Crosslinking studies suggest that the apo-E tetramer dissociates upon

lipoprotein binding, with the individual apo-E molecules floating freely on the particle surface (Funahashi, 1989). Whereas the linker region connecting the two domains is highly sensitive to proteases when apo-E is in solution, it becomes protease-resistant (suggesting a change in conformation or accessibility) upon lipoprotein binding (L.P. Aggerbeck, J. R. Wetterau, K.H. Weisgraber; personal communication). Circular dichroism shows no significant change in total α-helical content upon apo-E binding to dimyristoyl phosphatidylcholine, arguing against a major structural change within the stable domains upon lipid binding (Aggerbeck, 1988). It remains possible, however, that a rearrangement of the α-helices that preserves the total amount of α-helical structure (and is therefore spectroscopically undetectable) does occur. Apolipophorin-III, an insect apolipoprotein, has been hypothesized to unfold down its middle upon lipid binding, thereby exposing hydrophobic amino acids in its core for interaction with the non-polar lipid surface (Breiter, 1991). By engineering disulfides into the N-terminal domain of apo-E (which would hinder such a conformational rearrangement), it should be possible to test whether such gross structural changes are required for lipid binding.

Apolipoprotein receptors

Apo-E is one of the two apolipoproteins known to interact with the receptors specifically involved in lipoprotein clearance from the blood stream. Apo-B, the other receptor ligand and one of the largest human proteins sequenced to date ($M_r \approx 513$ kD), is quite hydrophobic and has been impossible to isolate in a stable lipid-free form. In contrast, apo-E is a fairly small protein ($M_r \approx 34$ kD) and is freely soluble in aqueous solutions to high concentrations. It is a striking paradox how these two proteins are both capable of binding the LDL receptor (also known as the apo-B,E receptor). While they show no significant sequence homology to each other, a single peptide of mostly basic amino acids is shared by both:

Apo-B:	3359	**Arg - Leu - Thr - Arg - Lys - Arg - Gly - Leu - Lys**	3367
		‖ ‖ ‖ ‖ ‖ ‖	
Apo-E:	142	**Arg - Lys - Leu - Arg - Lys - Arg - Leu - Leu - Arg**	150

A series of experiments, including neutralizing antibody epitope mapping (Weisgraber, 1983), site directed mutagenesis (Lalazer, 1988), and binding studies with proteolytic fragments (Innerarity, 1983), has shown that this region in apo-E is essential for LDL receptor binding. An intriguing possibility is that while the overall structures of apo-B and apo-E are fundamentally different, this conserved peptide may exist in a similar conformation in both proteins and thereby mediate LDL receptor binding. Because of the large size and insolubility of apo-B, it is unlikely that its structure will be determined crystallographically in the near future. If receptor binding is due to a common structural motif, the structure of apo-E (described in chapter two) may serve as a useful model for understanding the binding properties of both ligands.

Apo-E is known to bind with high affinity to two different receptors, the LDL receptor and the LDL receptor-related protein. The primary sequence of the LDL receptor (839 amino acids) indicates several distinct domains made up by multiple internal sequence repeats (Sudhof, 1985). At the N-terminus, seven copies of a ≈40-amino acid cysteine-rich motif are joined to form the apolipoprotein-binding domain. Three copies of an epidermal growth factor precursor motif following these class 'A' cysteine-rich repeats have been identified; their functional role remains unclear. Towards the C-terminus, a glycosylated linker region joins the repeats to a transmembrane domain (presumed to be a short α-helix), followed by a short cytoplasmic tail. The major function of the LDL receptor is to bind apo-B on the surface of LDL, thereby directing the endocytosis of particle. Once internalized, the lipoprotein particle is targetted for fusion with lyzosomes which degrade it and release its lipid and cholesterol for utilization by the cell. The number

of LDL receptors on the cell surface is tightly regulated to preserve cholesterol homeostasis. As cells become cholesterol-loaded, receptors are down regulated to decrease the influx of exogenous cholesterol. Increasing receptor numbers can have a dramatic effect on circulating cholesterol levels. Mevinolin, one of the most effective cholesterol-lowering drugs, acts by inhibiting the enzyme HMG-CoA reductase, directly leading to an increase in LDL receptor levels (Alberts, 1980). Increased receptor expression results in more efficient LDL clearance from the plasma, thereby lowering plasma cholesterol. Genetic defects in the LDL receptor are responsible for the well-characterized disorder familial hypercholesterolemia (FH) (Goldstein, 1989). Patients with this disease are unable to express functional LDL receptors and therefore accumulate large amounts of LDL, IDL, and VLDL remnants. LDL cholesterol levels in homozygous FH patients are dramatically higher than normal (700 mg/dl vs. 120 mg/dl), and as a result, patients almost always die of premature atherosclerosis.

As implied by its name, the LDL receptor-related protein (LRP) is a close homologue of the LDL receptor (Herz, 1988). This 600,000-kDa protein is synthesized largely in the liver and the brain. It contains 31 copies of the class-A cysteine-rich repeats and 22 copies of the EGF-precursor. Like the LDL receptor, LRP has high affinity for calcium. Whereas the LDL receptor binds both apo-B and apo-E with approximately equal affinity, LRP recognizes apo-E much more effectively than apo-B. It has been postulated that LRP may be the chylomicron remnant receptor since uptake of these particles by the liver is known to be apo-E-dependent. The true physiological role of this protein, however, remains to be experimentally determined.

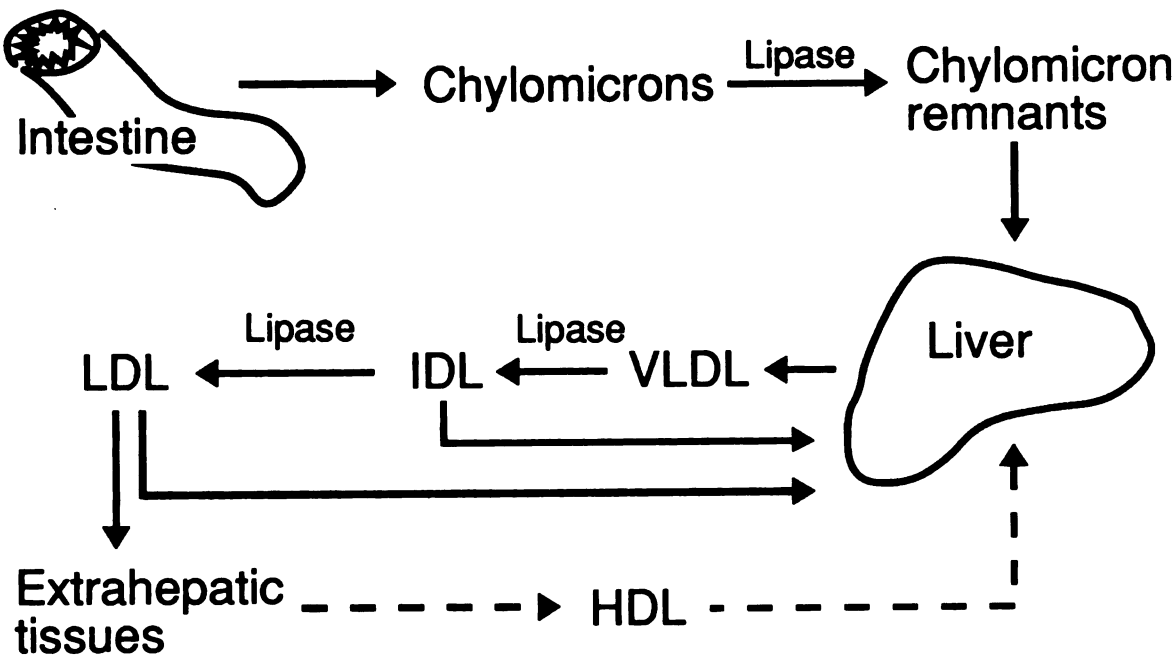


FIGURE 1.1. Pathways for lipid transport

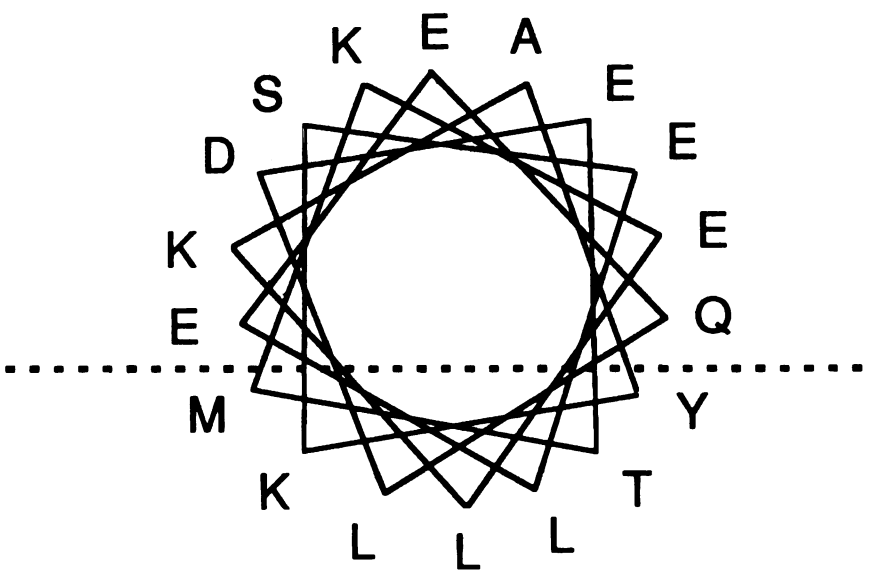


FIGURE 1.2. Helical wheel for the first 22-amino acid repeat of apo-E.

References

Aggerbeck, L.P., Wetterau, J.R., Weisgraber, K.H., Wu, C.-S.C., & Lindgren, F.T. (1988). Human apolipoprotein E3 in aqueous solution. II. Properties of the amino- and carboxyl-terminal domains. *J. Biol. Chem.* **263**, 6249-6258.

Alberts, A.W., Chen, J., Kuron, G., Hunt, V., Huff, J., Hoffman, C., Rothrock, J., Lopez, M., Joshua, H., Harris, E., Patchett, A., Monaghan, R., Currie, S., Stapley, E., Albers-Schonberg, G., Hensens, O., Hirschfield, J., Hoogsteen, K., Liesch, J., & Springer, J. (1980) Mevinolin, a highly potent competitive inhibitor of HMG-CoA reductase and cholesterol lowering agent. *Proc. Natl. Acad. Sci, USA.* **77**, 3957.

Beisiegel, U., Weber, W., Ihrke, G., Herz, J., & Stanley, K.K. (1989) The LDL-receptor-related protein, LRP, is an apolipoprotein E-binding protein. *Nature*, **341**, 162-164.

Breiter, D.R., Kanost, M.R., Benning, M.M., Wesenberg, G., Law, J.H., Wells, M.A., Rayment, I., & Holden, H.M. (1991). Molecular structure of an apolipoprotein determined at 2.5-Å resolution. *Biochemistry.* **30**, 603-608.

Brown, M.S. & Goldstein, J.L. (1983). Lipoprotein metabolism in the macrophage: implications for cholesterol deposition in atherosclerosis. *Annu. Rev. Biochem.* **52**, 223-261.

Funahashi, T., Yokoyama, S., & Yamamoto, A. (1989). Association of apolipoprotein E with the low density lipoprotein receptor: demonstration of its co-operativity on lipid microemulsion particles. *J. Biochem.* **105**, 582-587.

Goldstein, J.L., Brown, M.S. (1989). Familial hypercholesterolemia. In: Scriver, C.R., Beaudet, A.L., Sly, W.S., Valle, D., eds. *The metabolic basis of inherited*

disease. 6th ed. New York, N.Y., McGraw-Hill International Book Co., 1215-1250.

Herz, J., Hamann, U., Rogne, S., Myklebost, O., Gausepohl, H., & Stanley, K.K. (1988). Surface localization and high affinity for calcium of a 500-kD liver membrane protein closely related to the LDL-receptor suggest a physiological role as lipoprotein receptor. *EMBO Journal.* **7**, 4119-4127.

Hoshino, T., Yamashita, S., Sakai, N., Matsuzawa, Y., Tarui, S., & Kumasaka, K. (1989). A case of hyper-HDL-cholesterolemia presenting peculiar lipoprotein patterns in agarose gel electrophoresis. *Japanese Journal of Clinical Pathology*, **37**, 835-839.

Innerarity, T.L., Friedlander, E.J., Rall, S.C. Jr., Weisgraber, K.H., & Mahley, R.W. (1983). The receptor-binding domain of human apolipoprotein E. Binding of apolipoprotein E fragments. *J. Biol. Chem.* **258**, 12341-12347.

Kodama, T., Freeman, M., Rohrer, L., Zabrecky, J., Matsudaira, P., & Krieger, M. (1990). Type I macrophage scavenger receptor contains alpha-helical and collagen-like coiled coils. *Nature.* **343**, 531-535.

Knott, T.J., Pease, R.J., Powell, L.M., Wallis, S.C., Rall, S.C. Jr., Innerarity, T.L., Blackhart, B., Taylor, W.H., Marcel, Y., Milne, R., Johnson, D., Fuller, M., Lusis, J., McCarthy, B.J., Mahley, R.W., Levy-Wilson, B., & Scott, J. (1986). Complete protein sequence and identification of structural domains of human apolipoprotein B. *Nature*, **323**, 734-738.

Lalazar, A., Weisgraber, K.H., Rall, S.C. Jr., Giladi, H., Innerarity, T.L., Levanon, A.Z., Boyles, J.K., Amit, B., Gorecki, M., Mahley, R.W., & Vogel, T. (1988). Site-specific mutagenesis of human apolipoprotein E. Receptor binding activity of variants with single amino acid substitutions. *J. Biol. Chem.* **263**, 3542-3545.

Luo, C.-C., Li, W.-H., Moore, M.N., & Chan, L. (1986). Structure and evolution of the apolipoprotein multigene family. *J. Mol. Biol.* **187**, 325-340.

Mahley, R.W., Innerarity, T.L., Rall, S.C. Jr., Weisgraber, K.H. (1984). *J. Lipid Res.* **25**, 1277-???.

McLachlan, A.D. (1977) *Nature* **267**, 465-466.

Ruys, T., Strugess, I., Shaikh, M., Watts, G.F., Nordestgaard, B.G., & Lewis, B. (1989). *Lancet*, **11**, 1119-1122.

Sudhof, T.C., Goldstein, J.L., Brown, M.S., & Russel, D.W. (1985). The LDL receptor gene: A mosaic of exons shared with different proteins. *Science*. **228**, 815-822.

Weisgraber, K.H., Innerarity, T.L., Harder, K.J., Mahley, R.W., Milne, R.W., Marcel, Y.L., Sparrow, J.T. (1983). The receptor-binding domain of human apolipoprotein E. Monoclonal antibody inhibition of binding. *J. Biol. Chem.*, **258**, 12348-12354.

Wetterau, J.R., Aggerbeck, L.P., Rall, S.C. Jr., Weisgraber, K.H. (1988). Human apolipoprotein E3 in aqueous solution. I. Evidence for two structural domains. *J. Biol. Chem.*, **263**, 6240-6248.

Chapter 2:

The three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E

Abstract

Human apolipoprotein E, a blood plasma protein, mediates the transport and uptake of cholesterol and lipid via its high affinity interaction with different cellular receptors, including the low density lipoprotein (LDL) receptor. The three-dimensional structure of the LDL receptor-binding domain of apo-E has been determined at 2.5 angstrom resolution by x-ray crystallography. The protein forms an unusually elongated (65 Å) four-helix bundle, with the helices apparently stabilized by a tightly packed hydrophobic core that includes leucine zipper-type interactions and by numerous salt bridges on the mostly charged surface. Basic amino acids important for LDL receptor-binding are clustered into a surface patch on one long helix. This structure provides the basis for understanding the behavior of naturally occurring mutants that can lead to atherosclerosis.

Introduction

Lipids, including triglycerides, phospholipids, and cholesterol, are sparingly soluble in aqueous solution and are thus transported through the body in the form of lipoprotein particles. Apolipoproteins, the protein components of these particles, take part in stabilizing lipoproteins and directing their metabolism. Two apolipoproteins, apolipoprotein-E (apo-E) and apolipoprotein-B (apo-B), bind with high affinity to cell surface receptors, including the low density lipoprotein (LDL) receptor, and thereby mediate the cellular uptake of most lipoproteins, namely very low density lipoprotein (VLDL), low density lipoprotein (LDL), and high density lipoprotein (HDL) (1). Because plasma cholesterol concentrations and metabolism are unequivocally linked to the development of atherosclerosis and risk of coronary artery disease, it is of fundamental importance to characterize the interaction of apo-E and apo-B with the LDL receptor at the molecular level. As a first step, we have pursued high resolution crystallographic studies of apo-E.

Apo-E is a 299-residue protein (M_r = 34,200) that appears to be made up of two independently folded domains (*2*). Digestion with thrombin produces a 22-kD fragment (residues 1 to 191) corresponding to the NH_2-terminal domain and a 10-kD fragment (residues 216 to 299) corresponding to the COOH-terminal domain (Figure 2.1) (*2*). Extensive characterization of apo-E has revealed that the structural domains also define its functional domains. While the NH_2-terminal region is responsible for the binding of apo-E to the LDL receptor (*3*) and binds lipids only weakly (*4,5*), the COOH-terminal region mediates the binding of apo-E to the surface of lipoproteins (*4,5*) but does not bind to the LDL receptor (*3*).

Although the NH_2-terminal domain is related to other members of the apolipoprotein gene family (apo-AI, apo-AIV, apo-CI, apo-CII, apo-CIII), it is

fundamentally different in some respects. All these proteins contain internal sequence repeats predicted to form amphipathic alpha helices (*6,7*) and circular dichroism spectroscopy has shown that they all have a high helical content (*8*). However, whereas most apolipoproteins are only marginally stable and form aggregates in solution in the absence of lipid, the NH_2-terminal domain of apo-E exists as a monomer in solution at high concentrations and its free energy of stabilization ($\Delta G_{unfold} \approx 10$ kcal/mol) is typical of other globular soluble proteins (*2*). By contrast, the COOH-terminal domain of apo-E exhibits properties more typical of the other soluble apolipoproteins; it forms a tetramer in solution and displays a relatively low free energy of stabilization ($\Delta G_{unfold} \approx 3$ to 4 kcal/mol).

The properties of the NH_2-terminal domain of apo-E made it an attractive candidate for crystallographic analysis, and indeed it proved possible to grow high quality crystals (*9*). In contrast, the lack of suitable crystals for any other vertebrate apolipoprotein has made it impossible to perform high-resolution structural studies of these apolipoproteins. We report here the atomic structure of the 22-kD N-terminal thrombolytic fragment of human apo-E, based on the interpretation of a 2.5-Å electron density map.

Structure determination

Possibly as a result of slight changes in the protein purification procedure we used, we found it difficult to crystallize the NH_2-terminal domain following our previously reported procedure (*9*) and therefore searched for new crystallization conditions. Crystals suitable for x-ray diffraction studies were obtained with the 22-kD-thrombolytic fragment of human apo-E3 by the hanging drop method; we used 10 mg/ml protein, prepared as described in (*9*) with the addition of an ion exchange high performance liquid chromatography (HPLC) step, 15 percent PEG 400 (BDH), 20 mM sodium acetate-acetic acid buffer, 0.2 percent β-*n*-octyl-glucopyranoside (Calbiochem), and 0.1 percent β-

mercaptoethanol (final concentrations). Crystallization was induced by raising the pH of the crystallization reservoir from 4.5 to 5.3 (fully grown crystals were stable up to pH 7.2). The detergent β-*n*-octylglucopyranoside was not required for crystal growth but was helpful in reducing twinning. Using these conditions, we were able to routinely grow small perfect prisms. Precession photography showed that although the morphology of these crystals differed significantly from those reported (*9*), they belonged to the same space group ($P2_12_12_1$) and had approximately the same unit cell dimensions (a=41.3 Å, b=54.5 Å, c=87.0 Å, $\alpha=\beta=\gamma\equiv90°$).

Data were collected with a RIGAKU AFC5R diffractometer equipped with a graphite monochromator and a 600 mm detector arm. Early experiments indicated that the crystals were extremely susceptible to radiation decay. Crystals were rapidly frozen in a stream of boiling liquid nitrogen and data were collected at -160°C with an MSC low temperature apparatus (College Station, TX). Freezing resulted in a slight shrinkage of the unit cell (a=40.7 Å, b=54.0 Å, c=85.4 Å), as has been seen in most other low-temperature protein crystallography experiments (*10*). Freezing reduced radiation decay, making it possible to collect data for up to 2 weeks on a single crystal (Table 2.1).

A Patterson map was calculated from the native data to 4-Å resolution. The map contained a number of strong peaks (corresponding to overlapping interatomic vectors), which suggested the presence of a four-helix bundle. Standard molecular replacement techniques (*11*) with an idealized four-helix bundle as a search model indicated two possible configurations. Although rigid body refinement of either a four-helix bundle or a six-helix bundle (made up by two overlapping four-helix bundles) caused a reduction in the overall R-factor (10 to 4 Å data) to 44 to 48 percent (*12*), we were unable to identify portions of the molecule missing from the search model.

We screened approximately 40 different ionic heavy-atom compounds by diffractometry, none of which yielded a suitable isomorphous derivative. Although apo-E3 contains a single free cysteine (residue 112), none of the reactive mercurials tested were successful. However, equilibrating crystals with dimethyl mercury (a small, nonreactive, nonionic compound) did produce a useful derivative. Heavy atom refinement with the TCTREF program (*13*) indicated two high occupancy sites. Two long helices were easily identified in a map calculated with the heavy atom phases. Density modification (by solvent flattening) was carried out using the programs of Wang (*14*) and, while noise in the map was significantly reduced, the electron density remained too ambiguous to allow modeling of the structure (Figure 2.2A). To improve the phase estimates, we collected accurate anomalous data on the mercury derivative for those reflections predicted to have the strongest anomalous scattering signal (Table 2.1). Combined phase estimates together with density modification resulted in an easily interpretable map (Figure 2.2B).

A model containing residues 23 to 164 was constructed, and three cycles of X-PLOR (*15*) refinement and manual rebuilding reduced the crystallographic R factor for all data in the 8 to 2.5 Å range to 23.0 percent. The heavy atoms were found to occupy a hydrophobic pocket on the surface formed by Trp-34 and three neighboring leucines. The addition of 76 bound solvent molecules and residues 165-166 further reduced the R factor to 17.7 percent. Individual isotropic B-factors were refined for all atoms in the model. The electron density was calculated with phases based on the refined model and $2F_o$-F_c coefficients (Figure 2.2C). The rms deviation of bond lengths (0.017 Å) and bond angles (3.2°) for the refined model are typical of well-determined protein structures. All backbone dihedral angles (ϕ-ψ pairs) fall within allowed regions of the Ramachandran map, except for glycines at 127 and 165. Ninety-two percent of all side chains can be classified as rotamers in the Ponder and Richards library (*16*). Most of the non-rotamer side chains are

glutamate and methionine residues, which are poorly sampled in the Ponder and Richards data set.

Our final model lacks residues 1 to 22 and 167 to 191. At both ends of the modeled region, density is well determined but quickly disappears outside the model. Electrophoresis of the crystallized protein demonstrated that the complete 22-kD fragment was present, i.e. that the missing regions were not absent as a result of proteolysis. We believe residues 1 to 22 and 167 to 191 are disordered in the crystal and are not missing because of systematic phasing errors: (i) Digestion of apo-E with a battery of proteases shows that amino acids 1 to 20 and 165 to 191 are susceptible to proteolysis and thus likely to be unstructured (Figure 2.1) (*2*). (ii) The original SIR-SAS map, which is unbiased by the model, shows that there is no significant density remaining outside the modeled region. (iii) A difference map calculated between the bacterially expressed human protein and the native plasma-derived human protein fails to show a peak corresponding to the additional N-terminal methionine present in the bacterial protein, even at low resolution. (iv) The R factor is unlikely to drop to <18 percent with excellent stereochemistry if the missing residues contribute significantly to the scattering.

Apo-E: an unusually elongated four helix bundle

The NH_2-terminal domain of apo-E contains five helices making up >80 percent of the modeled residues (Figure 2.3). Four of the helices, containing 19, 28, 36, and 35 amino acids each, are arranged to form a 2 by 2 bundle. Four-helix bundles are the most common tertiary fold in α-helical proteins and are found in at least 18 other protein crystal structures (*17*). The average length for bundle helices in these structures is 18 residues, roughly half the helix length found for apo-E. Comparison to the core of myohemerythrin

(*18*), a typical four-helix bundle, indicates that the spacing between helices and the inter-helical angles are similar for the two proteins.

Each helix in the apo-E bundle lies antiparallel to the helices adjacent to it (this up-down topology is found in all other bundles except cytochrome P-450_{cam} (*17*)). The connection between the first two helices of the bundle (residues 24 to 42 and 54 to 81) is a short helix (residues 44 to 53). The turn between helix 2 and 3 of the bundle (residues 82 to 86) is poorly defined in the electron density map and in the refined model. The average crystallographic B factor for atoms in this loop is >60Å^2; completely omitting them causes an insignificant increase in the overall R factor. These residues have been included in the final model for completeness but may contain errors. Helix 3 (residues 87 to 122) is kinked at Gly 105 but maintains main chain hydrogen bonding through this region. Helix 4 (residues 130 to 164), containing residues known to be important in LDL receptor binding, is well ordered throughout.

Internal sequence repeats and apolipoprotein evolution

Apo-E shares a gene structure common to most other apolipoproteins (apo-AI, apo-AIV, apo-CI, apo-CII, and apo-CIII) (*6,7,19*). In all of these proteins, the coding region of the mature peptide is split between exons three and four. Exon 3 includes three consecutive 11-amino acid repeats while exon 4 contains 1 to 12 copies of a 22-amino acid repeat (*6,7*). The protein sequence identity between the apo-E repeats is relatively low compared with that in other apolipoproteins and is barely detectable in the nucleotide coding sequence. Figure 2.3 shows a clear structural role for these repeats as the basis for the long helices in the bundle. Helix 1 and the connecting helix are defined by the first two 11-aa repeats of exon 3. Helix 2 contains the third 11-aa repeat and one 22-aa repeat, while helices 3 and 4 are each made up of two 22-aa repeats. The apparently disordered COOH-terminal residues (167 to 191) correspond to parts of the fifth and sixth 22-aa repeats.

Multiple copies of short internal repeats have been detected in several different protein sequences (for example, the insect protein apolipophorin-III (*20*), the leucine-rich α_2-glycoprotein protein of human serum (*21*), and the Tau and MAP2 microtubule-binding proteins (*22, 23*)). The crystal structure of apo-E provides clear evidence that repeats such as these can function as structural building blocks within a single domain.

The boundaries between internal repeats appear to have two alternative functions in the folded structure. At residues 40, 51, 84, and 128, the junctions form turns, placing neighboring repeats into separate helices. At residues 62, 106, and 150, the junctions adopt a helical conformation and maintain the two adjacent repeats in a single helix. The helix built up by the third and fourth 22-amino acid repeats (helix 3) is strongly kinked at the repeat junction (residues 105 to 106). None of the turns in the structure correspond to the middle of a repeat. This suggests that, while the ends of the repeats are not limited to a unique conformation, the cores of the repeats are inherently helical. In a simple model for the structure and evolution of apolipoproteins, the internal repeats correspond to stable helices that can pack well with their helical axes antiparallel to one another. Whether adjacent sequence repeats exist in the same or in separate helices is likely to be determined by the intervening junction sequences. An understanding of the sequence-dependence of stability is needed to predict the topology of other apolipoproteins.

As predicted from analysis of apolipoprotein sequences (*24*), the helices of apo-E are strongly amphipathic. The four long helices of the bundle are arranged such that their hydrophobic residues are completely sequestered inside the protein while their hydrophilic faces are solvent exposed (Figure 2.4). Leucine side chains occurring about every seven residues appear to stabilize the interfaces between helix 1 and helix 4 and between helix 2 and helix 3 (Figure 2.5). The organization of leucine interactions on adjacent antiparallel helices is very similar to the leucine-zipper model of Landschulz *et al.* (*25*), predicted for

the dimerization domain of C/EBP-type transcription factors. Although subsequent experimental work has shown that the Landschulz model is probably not correct for the transcription factors (*26*), the crystal structure of apo-E indicates that the original leucine-zipper model is an energetically feasible way of stabilizing pairs of helices.

More than a third of the NH_2-terminal domain residues are charged. The crystal structure includes atomic coordinates for 24 acidic and 24 basic residues. These amino acids almost completely cover the surface of the bundle (Figure 2.4), and most participate in either intramolecular or intermolecular salt bridges in the crystal structure. Of the intramolecular salt bridges, eleven are formed by pairs of amino acids lying in the same helix. Surprisingly, more than half of these salt bridges are oriented to interact unfavorably with the macrodipole of the helix to which they belong. Seven salt bridges are formed between pairs of helices in the bundle. These strong electrostatic interactions may help to bind the helices together and further stabilize the folded structure. The combination of tight hydrophobic packing and numerous electrostatic interactions would seem to account for the high free energy of stabilization for this domain, setting it apart from most other apolipoproteins.

An electrostatic potential map was calculated for the 22-kD fragment with the DELPHI program (Figure 2.6) (*27*). The DELPHI algorithm solves the Poisson-Boltzmann equation with the use of the formal charge distribution indicated by the crystal structure. Because most charged residues are paired to form salt bridges, the net electrostatic potential is close to zero for most of the region surrounding the protein. The only significant feature of this map is a large region of positive potential encompassing the NH_2-terminal half of helix four (residues 136 to 150). The possible role of this electrostatic feature in LDL receptor binding is discussed below.

Basis for LDL receptor binding and lipid binding

One of apo-E's major functions is in mediating the cellular uptake of lipoprotein particles (for review, see ref. 1). Lipoprotein-associated apo-E binds with high affinity to LDL receptors on the surface of target cells. Upon apolipoprotein binding, these receptors cluster in clathrin-coated membrane pits, which subsequently pinch off to internalize the lipoprotein particle. This process eventually results in the degradation of the lipoprotein particle and the release of its lipids for cellular use. Apo-E also functions in the uptake of chylomicron remnants by liver cells via an apo-E-specific receptor - possibly LRP, the LDL-receptor related protein (*28*).

Low density lipoprotein receptor-binding is localized in the NH_2-terminal domain of apo-E (*3*). Genetic analysis of point mutations in apo-E, in combination with site-directed mutagenesis and antibody competition studies, have shown that residues 136 to 150 are necessary for this function (*3,29*). Electron density for all of these residues, found to lie on helix 4 of the bundle, is well determined. The NH_2-terminal portion of helix 4 is unusually rich in basic residues, including Arg 134, Arg 136, His 140, Arg 142, Lys 143, Arg 145, Lys 146, Arg 147, and Arg 150. Because most of these amino acids are solvent-exposed and not involved in intramolecular salt bridges, their excess positive charge combines to produce a large region of positive electrostatic potential extending 15 Å out of the protein (Figure 2.6). Naturally occurring variants are known for positions 136, 142, 145, and 146 in which the basic amino acid is substituted by a neutral (*3,25*) or in one case by an acidic amino acid (*30*). All of these variants display defective binding to the LDL receptor (20 to 40 percent of normal binding). Site-directed mutagenesis of other basic residues in the helix 4 cluster, Lys 143 or Arg 150, also reduces receptor binding to 10 to 50 percent of normal levels (*31*). The fact that no single substitution results in complete abolition of binding activity suggests that there are multiple interactions between apo-E and

the receptor and that the basic residues may cooperate in the binding function. That these basic residues are not involved in salt bridges within the structure supports the hypothesis that they are free to interact with the receptor.

The LDL receptor recognizes both apo-E and apo-B, with comparable affinity (*32*). While there is essentially no sequence similarity between these two ligands, apo-B contains a region rich in basic residues that has a similar pattern of charged and neutral amino acids as that in the receptor-binding helix of apo-E. Several studies have indicated that this region may play a role in receptor binding (*33*). It is possible that a positively charged helix is used by both proteins to specify LDL receptor binding. Because of apo-B's insolubility and large size (>500-kD), it is unlikely that the structure of this medically-relevant protein will be determined crystallographically in the near future. As such, the structure of apo-E's receptor-binding helix may provide a useful starting point for modelling the interaction of this other important ligand with the LDL receptor.

Although the binding of apo-E to lipoprotein particles appears to be mediated by the C-terminal domain (residues 216 to 299) (*4,5*), the NH_2-terminal domain does associate *in vitro* with phospholipid to form discoidal particles (*3*). In the lipid-bound form apo-E is able to bind with very high affinity to the LDL receptor ($K_d \approx 1.2 \times 10^{-10}$ M) (*34*) (recent work suggests that the lipid-free protein binds with ≈500-fold lower affinity than the lipid-complexed protein (*35*)). Several studies had suggested that helices in the N-terminal domain (especially helix 4) may insert into phospholipid membranes. In support of this, short synthetic peptides corresponding to helix 4 have been shown to have lipid binding activity (*36*). By contrast, the crystal structure does not indicate a special lipid binding role for helix 4 or any of the other helices. Hydrophobic amino acids in the 136 to 150 region are all buried and well packed, interacting with the hydrophobic residues of helices 1 and 3. In addition, the α-helical hydrophobic moments for all of the core helices fall in the same

range as those for the helices in other four helix bundle proteins, none of which have significant lipid binding activity.

The lack of hydrophobic patches on the protein surface and the stability of the globular structure suggest several other possible explanations for lipid binding. For example, lipid binding might involve a significant conformational change in which the hydrophobic face of one or more helices become accessible for interaction with non-polar phospholipid tails. Alternatively, lipid binding could result from the interaction of charged surface amino acids and phospholipid head groups. Finally, it is possible that the missing disordered residues (1 to 22, 167 to 191) become structured in the presence of lipid and mediate lipid binding. Previous biophysical studies have indicated that the 22-kD fragment of apo-E does not undergo a detectable conformational change on lipid binding (*2*); a result that does not support the first hypothesis. Furthermore, we have found that the linker between the NH_2- and COOH-terminal domains becomes less protease-sensitive upon lipid binding (*37*). Although these data support a model in which lipids interact with the COOH-terminal region of the 22-kD domain and orders it, further crystallographic and spectroscopic studies are required to elucidate the mechanism of lipid binding and its role in activating the binding of apo-E to the LDL receptor.

Genetic variability in apo-E

Three major isoforms of apo-E have been characterized; these are apo-E2, E3, and E4. Functional differences between these proteins appear to lead to quantitative changes in both plasma cholesterol and in the likelihood of coronary heart disease (reviewed by Davignon *et al.*, ref. *38*). Apolipoprotein-E3, the protein used for our study, is the most common isoform. The replacement of arginine by cysteine at position 158 (corresponding to the most common apo-E2 isoform) results in defective LDL receptor binding (*39*) and is strongly linked to type III hyperlipoproteinemia, a genetic disorder associated with

premature atherosclerosis. The mutation of Cys 112 to Arg (apo-E4) is associated with increased plasma and LDL cholesterol levels, despite normal LDL receptor binding (*39*). In the absence of refined structures for these mutant proteins, we can speculate on the structural basis for their functional defects.

Arginine 158, site of the E2 mutation, is located near the COOH-terminal end of helix 4, well removed (>15 Å) from the other residues implicated in LDL receptor binding. The guanidinium group of Arg 158 does not contribute directly to the large positive electrostatic potential surrounding the receptor-binding helix. Instead, it forms salt bridges with the acidic side chains of Glu 96 and Asp 154, and as such may help to stabilize the pairing of helices 3 and 4. Previous experiments have suggested that the E2 mutation may induce a significant conformational change that reduces receptor binding (*40*). The finding that residue 158 lies far from the basic region of residues 136-150 yet has a dramatic effect on receptor binding is consistent with this hypothesis.

The Cys 112 to Arg mutant (the apo-E4 isoform) has full LDL receptor binding (*39*) but altered lipoprotein binding (strongly favoring VLDL over HDL) (*41*). This effect is surprising since the mutation occurs in the NH_2-terminal domain, yet lipoprotein binding is largely mediated by the C-terminal domain. Cys 112 is partially buried between helices 2 and 3, well isolated from the receptor-binding helix (Figure 2.2C). Model building based on the apo-E3 structure suggests that an arginine at this position could be easily accommodated by filling the solvent region surrounding the helix pair. In doing so, the arginine side chain may disrupt a specific interaction between the NH_2-terminal domain and the COOH-terminal domain, altering the structure or accessibility of the lipid binding determinants directly. Further crystallographic studies of the E2 and E4 isoforms can test these hypotheses.

NOTE ADDED IN PROOF. Crystallographic studies of the *Locust migratoria* protein apolipophorin III have been reported since the submission of our manuscript (*42*). The structure of this protein is similar to that for apo-E, suggesting that the elongated helical bundle may be a common structural motif for apolipoproteins. We have recently extended the resolution of the apo-E data to 2.25 Å; coordinates for the newly refined structure will be deposited with the Protein Data Bank (Brookhaven), code 1LPE.

TABLE 2.1 Statistics for crystallographic data and refinement. Data were collected at liquid nitrogen temperatures with a Rigaku AFC5R diffractometer (using a rotating anode generator operated at 180 mA, 50 kV). The native crystals belong to space group $P2_12_12_1$ and had unit cell dimensions a=40.65 Å, b=53.96 Å, c=85.43 Å. Native data represent the merged observations from three crystals. The isomorphous derivative data was obtained by diffusing dimethyl mercury into fully grown crystals for 2 weeks. Isomorphous data were collected from two crystals and anomalous data were collected from a single crystal (statistics for anomalous data are shown in parentheses).

	Native	$(CH_3)_2Hg$ Derivative	
Resolution (Å)	2.5	2.5	(2.8)
Diffraction data			
No. of observations	17129	9957	(3317)
No. of unique reflections	6899	6880	(1499)
Completeness (percent)	100	100	(31)
$\langle I \rangle / \langle \sigma_I \rangle$	43.48	30.42	(51.58)
R_{merge} *	0.054		
Phasing statistics			
Heavy atom sites		2	
Isomorphous differences		6474	
Anomolous differences			(1244)
r.m.s. F_H/residual †		1.55	(1.57)
R_{Cullis} ≠		0.55	(0.50)
r.m.s. R_{anom} ≠			(0.48)
<figure of merit>		0.41	(0.53)
<figure of merit following solvent flattening>		0.56	
X-PLOR refinement statistics			
Number of protein atoms (non-hydrogens)	1172		
Number of water molecules	76		
R_{cryst} (overall)	0.177		
8.50 Å - 3.85 Å	0.149		
3.85 Å - 3.12 Å	0.171		
3.12 Å - 2.74 Å	0.204		
2.74 Å - 2.50 Å	0.223		
r.m.s. deviations from ideality			
bond lengths (Å)	0.017		
bond angles (°)	3.2		

*$R_{merge} = \frac{\Sigma\Sigma \mid I_j(h) - \langle I(h) \rangle \mid}{\Sigma\Sigma I_j(h)}$ (summations done over all reflections in all crystals).

† r.m.s. F_H/residual = phasing power where F_H = r.m.s. heavy atom structure factor amplitude (calculated) and residual = r.m.s. lack of closure error.

≠ $R_{Cullis} = \frac{\Sigma \mid FPH_{calc} - FPH_{obs} \mid}{\Sigma \mid FPH_{obs} - FP \mid}$ (for centric reflections only, FPH and FP are structure factors for the derivative and native protein respectively).

≠ r.m.s. R_{anom} = r.m.s. phase-averaged anomalous residual divided by the r.m.s. observed anomalous difference. $R_{cryst} = \frac{\Sigma \mid F_{obs} - F_{calc} \mid}{\Sigma \mid F_{obs} \mid}$.

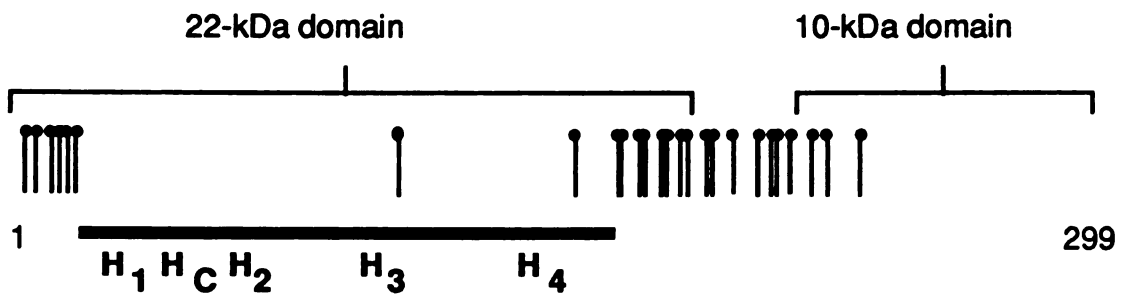


FIGURE 2.1. Domain structure of apo-E. Residues 1-191 define the N-terminal domain of apo-E, known to be important for LDL receptor binding. Solid circles indicate proteolytically sensitive residues (as determined by digestion with thrombin, elastase, trypsin, chymotrypsin, S. aureus V8, and subtilisin (*2*)). The solid bar indicates the region that is included in the refined crystal structure of the 22-kDa fragment (residues 24-166). The helices (shaded areas) making up the four-helix bundle are numbered as follows: H_1, residues 24-42; H_2, 54-81; H_3, 87-122; and H_4, 130-164. A connecting helix (H_c, residues 44-53) joins helices 1 and 2. Secondary structure assignments were made using the method of Kundrot and Richards (*35*). The complete apo-E sequence includes three very weakly conserved 11-mers (29-39,40-50,51-61) and eight 22-mers (62-83,84-105,106-127,128-149,150-171,172-193,194-215,216-237) (*19*).

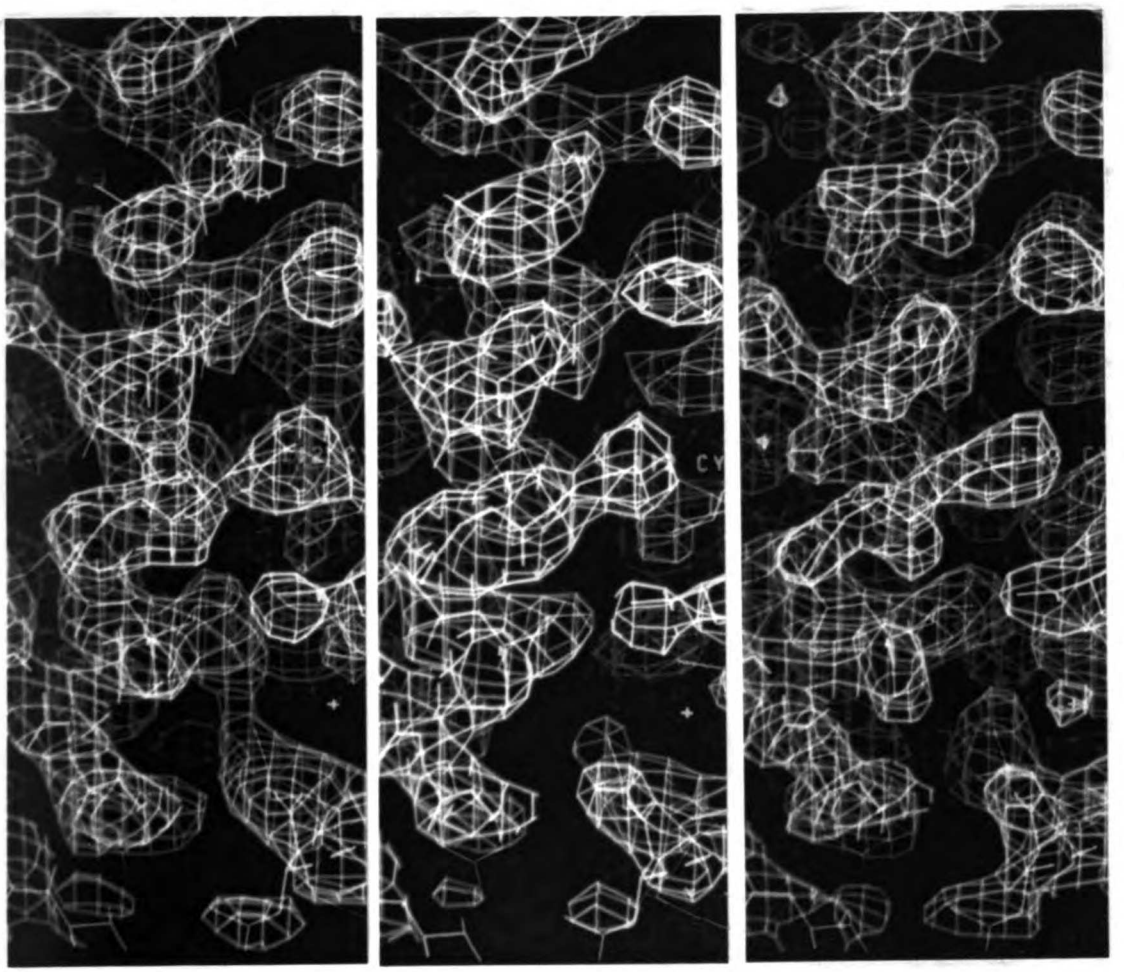


FIGURE 2.2. Electron density calculated at various stages of phase refinement. All pictures show the region surrounding residue Cys 112, site of the mutation in the apo-E4 isoform. a) Electron density map calculated using SIR phases from the dimethyl mercury derivative, following extensive solvent flattening. b) Map calculated using SIR + anomalous phases, also following solvent flattening. c) $2F_o$-F_c map calculated using the final α_c phases.

FIGURE 2.3. Ribbon diagram of the N-terminal domain based on the crystal structure. Internal sequence repeats are used to color code the α-helices. The three 11-residue repeats are colored blue and cyan, while the five 22-residue repeats are colored red and yellow.

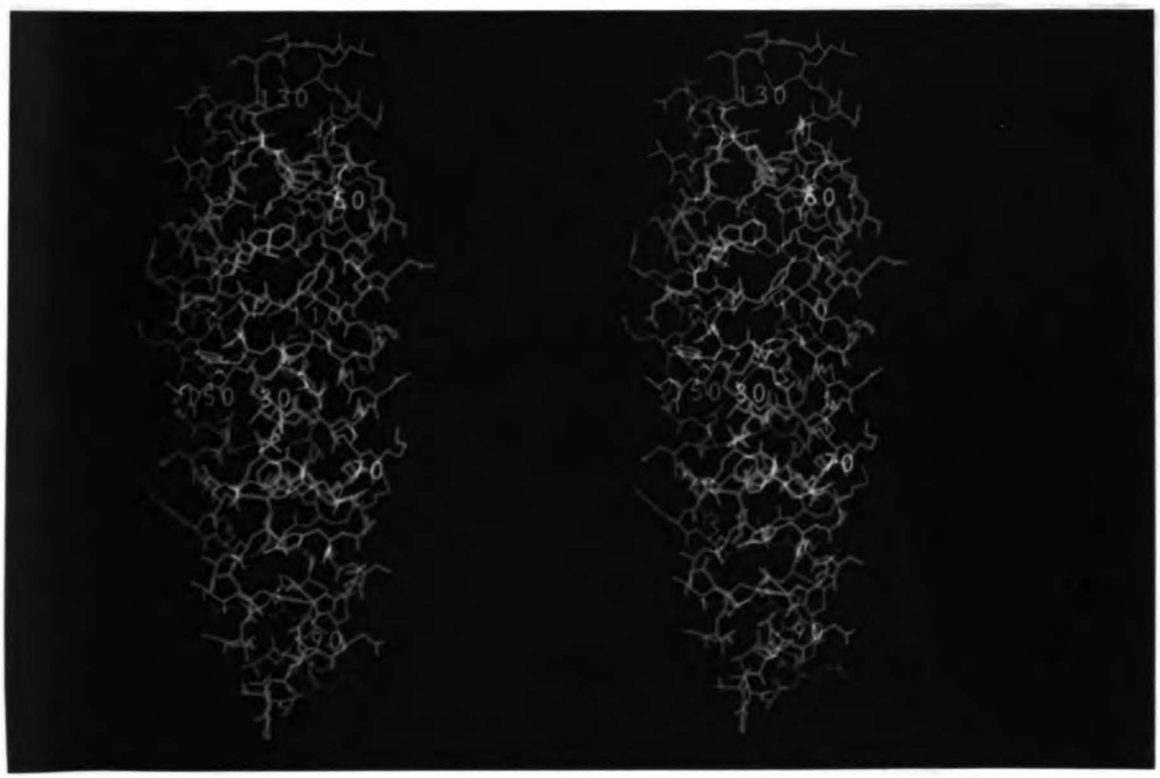


FIGURE 2.4. Stereo view of the refined atomic model of the LDL receptor-binding domain. Hydrophobic residues (green) line the buried faces of the bundle helices and form a well-packed core. Charged amino acids (red) cover most of the surface.

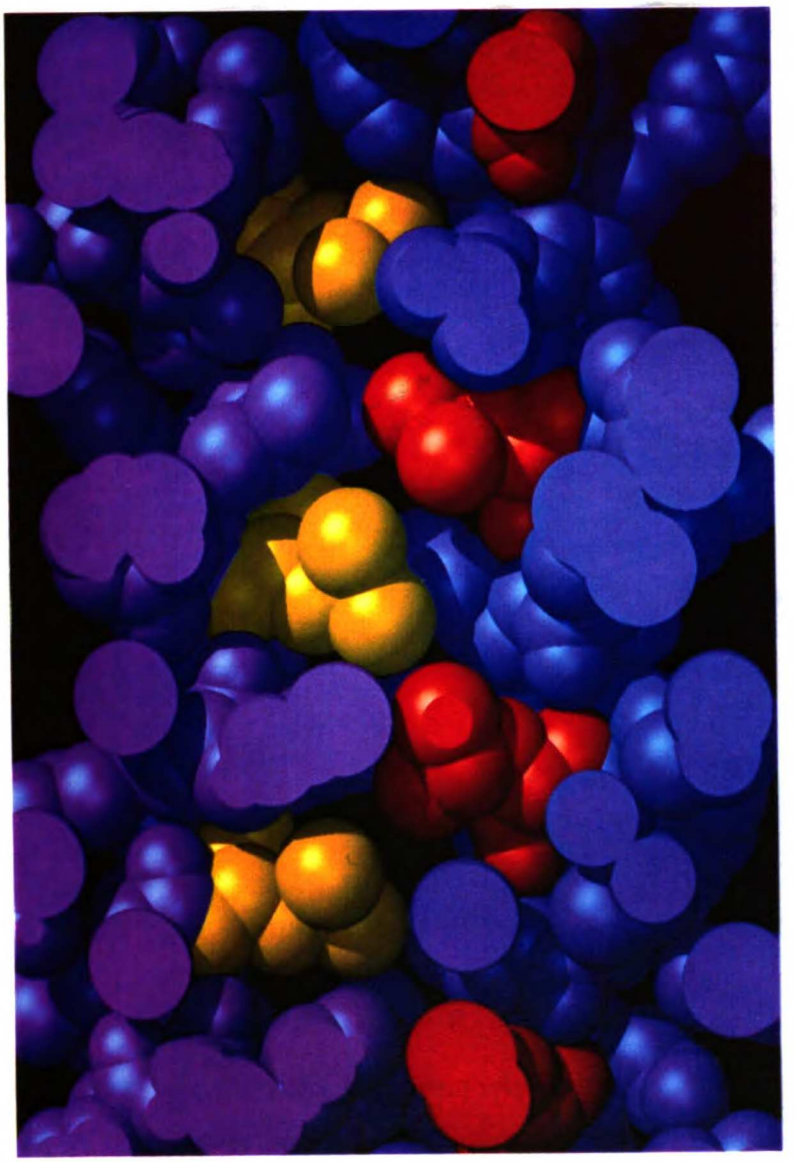


FIGURE 2.5. Cross-section through a space-filling representation of the N-terminal domain. Leucines on helix 1 (30,37,43; yellow) and helix 4 (133,141,148,155; red) are arranged to form a "zipper" in a manner similar to that predicted by Landschulz *et al.* (*21*) for the C/EBP-type transcription factors.

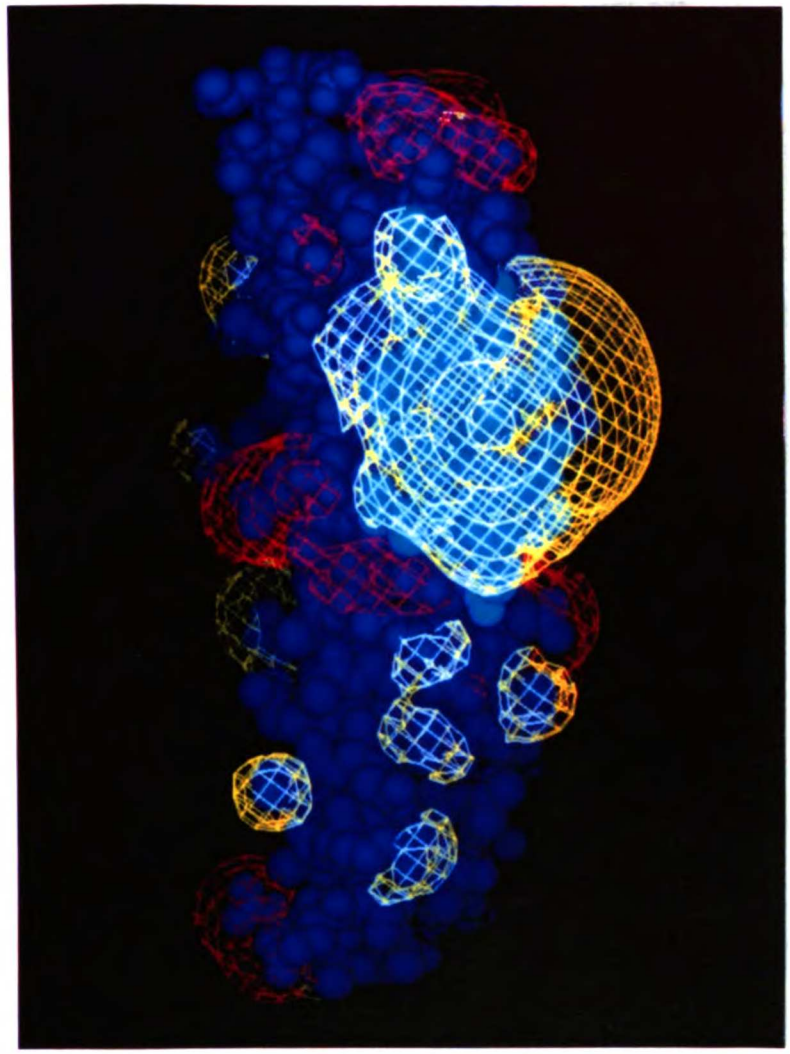


FIGURE 2.6. Electrostatic potential map of apo-E. The DELPHI program (Biosym, San Diego, CA) was used to calculate an approximate solution to the linearized Poisson-Boltzmann equation. The dielectric of the protein region was set to 2, while the solvent dielectric was set to 80. An ionic strength of 150 mM was assumed and only formal protein charges were included in the calculation. Positive (yellow) and negative (red) contours in the potential are evaluated at +2 and -2 kT/e^- respectively and are shown together with a space-filling model of the entire N-terminal domain (blue). Residues in the receptor-binding region (136-150) are colored cyan.

References

1. R. W. Mahley, *Science* **240**, 622 (1988).

2. J. R. Wetterau, L. P. Aggerbeck, S. C. Rall, Jr., K. H. Weisgraber, *J. Biol. Chem.* **263**, 6240 (1988); L. P. Aggerbeck, J. R. Wetterau, K. H. Weisgraber, C.-S. C. Wu, F. T. Lindgren, *J. Biol. Chem.* **263**, 6249 (1988).

3. T. L. Innerarity, E.J. Friedlander, S. C. Rall, Jr., K. H. Weisgraber, R. W. Mahley, *J. Biol. Chem.* **258**, 12341 (1983).

4. S. H. Gianturco, A. M. Gotto, Jr., S.-L. C. Hwang, J. B. Karlin, A. H. Y. Lin, S. C. Prasad, W. A. Bradley, *J. Biol. Chem.* **258**, 4526 (1983).

5. K. H. Weisgraber, *J. Lipid Res.* **31**, 1503 (1990).

6. M. S. Boguski, M. Freeman, N. A. Elshourbagy, J. M. Taylor, J. I. Gordon, *J. Lipid Res.* **27**, 1011 (1986). M. S. Boguski, E. H. Birkenmeirer, N. A. Elshourbagy, J. M. Taylor, J. I. Gordon, *J. Biol. Chem.* **261**, 6398 (1986).

7. C.-C. Luo, W.-H. Li, M. N. Moore, L. Chan, *J. Mol. Biol.* **187**, 325 (1986).

8. H. J. Pownall, Q. Pao, D. Hickson, J. T. Sparrow, A. M. Gotto, Jr. *Biophys. J.* **37**, 175 (1982).

9. L. P. Aggerbeck, J. R. Wetterau, K. H. Weisgraber, R. W. Mahley, D. A. Agard *J. Mol. Biol.* **202**, 179 (1988).

10. H. Hope, *Ann. Rev. Biophys. & Biophys. Chem.* **19**, 197 (1990).

11. R. A. Crowther. In *Molecular Replacement Methods* (ed M.G. Rossmann), Gordon & Breach, N.Y. (1971).

12. Refinement done using the CORELS program (J. L. Sussman, *Meth Enzym.* **115**, 271 (1985)).

13. Refined using the HEAVY program (T. C. Terwilliger, D. Eisenberg, *Acta Cryst* **A39**, 813-817 (1983)), modified by V. Ramalingam, UCSF.

14. B. C. Wang, *Meth. Enzym.* **115**, 90 (1985).

15. A. T. Brunger, J. Kuriyan, M. Karplus, *Science* **235**, 458 (1987).

16. J. W. Ponder and F.M. Richards, *J. Mol. Biol.* **193**, 775 (1987).

17. S. R. Presnell and F.E. Cohen, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6592 (1989).

18. S. Sheriff, W. A. Hendrickson, J. L. Smith, *J. Mol. Biol.* **197**, 273 (1987).

19. H. K. Das, J. McPherson, G. A. P. Bruns, S. K. Karathanasis, J. L. Breslow, *J. Biol. Chem.* **260**, 6240 (1985).

20. M. R. Kanost, M. S. Boguski, M. Freeman, J. I. Gordon, G. R. Wyatt, M. A. Wells, *J. Biol. Chem.* **263**, 10568 (1988).

21. N. Takahashi, Y. Takahashi, F. W. Putnam, *Proc. Natl. Acad. Sci. USA.* **82**, 1906 (1985).

22. G. Lee, N. Cowan, M. Kirschner, *Science*, **239**, 285 (1988).

23. S. A. Lewis, D. Wang, N. J. Cowan, *Science*, **242**, 936 (1988).

24. J. P. Segrest, R. L. Jackson, J. D. Morrisett, A. M. Gotto, Jr., *FEBS Lett.* **38**, 247 (1974).

25. W. H. Landschulz, P. F. Johnson, S. L. McKnight, *Science* **240**, 1759 (1988).

26. E. K. O'Shea, R. Rutkowski, P. S. Kim, *Science* **243**, 538 (1989).

27. M. K. Gilson and B. H. Honig, *Nature* **330**, 84 (1987).

28. U. Beisiegel, W. Weber, G. Ihrke, J. Herz, K. K. Stanley, *Nature* **341**, 162 (1989).

29. K. H. Weisgraber *et al., J. Biol. Chem.* **258**, 12348 (1983).

30. W. A. Mann, R. E. Gregg, D. L. Sprecher, H. B. Brewer, Jr., *Biochim Biophys Acta* **1005**, 239 (1989).

31. A. Lalazar *et al., J. Biol. Chem.* **263**, 3542 (1988).

32. R. E. Pitas, T. L. Innerarity, R. W. Mahley, *J. Biol. Chem.* **255**, 5454 (1980).

33. T. J. Knott *et al.*, *Nature* **323**, 734 (1986); C.-Y. Yang *et al.*, *Nature* **323**, 738 (1986).

34. T. L. Innerarity, R. E. Pitas, R. W. Mahley, *J. Biol. Chem.* **254**, 4186 (1979).

35. J. R. Wetterau, L. P. Aggerbeck, K. H. Weisgraber, unpublished observations.

36. J. T. Sparrow, D. A. Sparrow, A. R. Culwell, A. M. Gotto, Jr., *Biochemistry* **24**, 6984 (1985).

37. K. H. Weisgraber, unpublished observations.

38. J. Davignon, R. E. Gregg, C. F. Sing, *Arteriosclerosis,* **8**, 1-21 (1988).

39. K. H. Weisgraber, T. L. Innerarity, R. W. Mahley, *J. Biol. Chem.* **257**, 2518 (1982).

40. T. L. Innerarity, K. H. Weisgraber, K. S. Arnold, S. C. Rall, Jr., R. W. Mahley, *J. Biol. Chem.* **259**, 7261 (1984).

41. R. E. Gregg *et al.*, *J. Clin. Invest.* **78**, 815 (1986).

42. D. R. Breiter *et al.*, *Biochemistry* **30**, 603 (1991).

39. C.E. Kundrot and F.M. Richards, *Proteins* **3**, 71 (1988).

Chapter 3 :

Structural basis for defective function in common mutants of human apolipoprotein-E

Apolipoprotein-E (apo-E), a 34-kDa blood plasma protein, plays a key role in directing cholesterol and lipid transport through the body. The apo-E gene is polymorphic, and two common mutant alleles (ε2 and ε4), corresponding to single-site substitutions in the LDL receptor-binding domain, are associated with significant changes in plasma cholesterol levels and in the risk for coronary heart disease[1]. We have recently determined the three-dimensional structure of the apo-E receptor-binding domain via x-ray crystallography[2]. By directly examining the structural changes that occur for each of these two common mutants, we hoped to understand the basis for altered apo-E function. We now report the refined crystal structures of the N-terminal domain of apo-E2 and apo-E4, determined to 3.0 Å and 2.5 Å resolution respectively. The apo-E2 structure reveals significant conformational rearrangements that can account for reduced LDL-receptor binding. No major changes are observed in the apo-E4 structure, suggesting that this mutation may alter lipoprotein binding by indirect means (for example, by modifying the interaction between the amino- and carboxy-terminal domains of apo-E).

The high affinity binding of apo-E by cell-surface receptors, including the low density lipoprotein (LDL) receptor, allows lipoproteins associated with apo-E to be targeted for endocytosis and intracellular degradation[3]. Interference with such receptor-mediated processing can cause lipoproteins to accumulate in the plasma and can ultimately lead to the formation of atherosclerotic plaques[1]. The two common point mutants of apo-E were initially identified by their altered electrophoretic mobility and are termed apo-E2 and apo-E4[4]. Relative to apo-E3 (the wild-type protein), the most common apo-E2 isoform is characterized by the substitution Arg-158→Cys[5]. This mutation (allelic frequency ≈0.08) lowers LDL receptor binding to <2% of normal levels, although the protein appears to bind to lipoproteins with the same affinity and specificity as the wild-type protein[6]. In contrast, the apo-E4 isoform (allelic frequency ≈0.15), corresponding to substitution Cys-112→

Arg, has normal LDL receptor-binding but markedly altered lipoprotein binding properties[7]. While plasma cholesterol and LDL concentrations are lowered in people expressing the apo-E2 protein, levels are raised in apo-E4 individuals[8]. Epidemiological studies suggest that apo-E2 has a protective function against coronary heart disease (CHD), although a small fraction of apo-E2 homozygotes are predisposed to type III hyperlipidemia, a form of atherosclerosis[9]. Ethnic groups enriched for the ε4 allele (*e.g.* Finns) are also at high risk for CHD, suggesting that the LDL-raising properties of this mutant may be significant in terms of altering the risk for atherosclerosis[10].

Apo-E appears to be organized as a 22-kDa N-terminal LDL receptor-binding domain and a 10-kDa C-terminal lipoprotein-binding domain, separated by a short 2-kDa linker[11]. We have previously shown that the LDL receptor-binding domain is an unusually elongated four-helix bundle[2]. Figure 3.1 shows a ribbon diagram of this domain and indicates the sites of the E2 and E4 mutations. Surprisingly, both substitutions are physically well removed from the cluster of basic residues known to be important for LDL receptor binding (136-150). The 22-kDa thrombolytic fragment of the apo-E mutants was isolated as described previously using blood plasma from human donors[12]. Details for the crystallization, data collection, and structure refinement for both mutants are summarized in Table 3.1.

The apo-E2 mutation (Arg-158→Cys) results in significant structural rearrangements in a large zone around the substituted residue. The rms deviation between equivalent atoms in the wild-type and mutant structures within 10 Å of the mutation site is 1.6 Å, significantly higher than that for all atoms (<1.0 Å). In the wild-type protein, the side chain carboxylates of Glu-96 and Asp-154 are paired with the oppositely-charged guanidinium group of Arg-158. These salt bridges form part of an extended network of charge interactions that link together helix 3 (87-123) and helix 4 (130-164) of the bundle (Figure 3.2a). When the positively-charged Arg-158 is substituted by a neutral cysteine,

this pattern of native salt bridges is completely disrupted. Glu-96 and Asp-154 move away from their initial positions and pair with a new set of positively-charged residues. To accommodate this rearrangement, other salt bridges are broken and additional pairs form (Figure 3.2c). The original set of seven salt bridges is replaced by only two new salt bridges.

Significant distortions in the peptide backbone of the apo-E2 protein arise as a result of the mutation. The N-terminal half of helix 3 (residues 87-104) is displaced away from the remainder of the bundle (Figure 3.2e), and a slight kink in the helix at Gly-105 becomes exaggerated in the mutant protein. This displacement of the helix is accompanied by a significant disruption of the α-helical geometry of the protein backbone. While the same pattern of hydrogen-bonding is maintained, several hydrogen-bonds are stretched and the (ϕ,ψ) angles of these residues no longer lie in the low-energy region of the Ramachandran map.

While lacking a structure of the protein complexed to the LDL receptor, we can speculate on the mechanisms by which receptor binding is disrupted for the E2 mutant. Site-directed mutagenesis has shown that a number of basic residues in the region 136-150 are required for full receptor binding. Replacement of one of these key residues, Arg-150, by alanine reduces LDL-receptor binding to one-quarter of normal values. In the apo-E3 structure, Arg-150 is solvent-exposed, presumably accessible for interaction with complementary acidic residues on the LDL receptor. In the E2 structure, however, this residue has swung out of the highly-positive region of helix 4 and is paired with Asp-154 (Arg-150 thus serves as an alternate for the replaced Arg-158). This rearrangement significantly alters the electrostatic potential surrounding the receptor-binding helix (Figure 3.2b,d). The combination of significant electrostatic and steric changes in the receptor-binding region would seem sufficient to fully account for the reduced binding of this mutant.

In marked contrast to the apo-E2 mutation, the apo-E4 mutation (Cys-112→Arg) causes very slight structural changes which are well localized to the site of the substitution (Figure 3.3). The r.m.s. deviation of backbone atoms between the E3 and E4 structures (excluding the poorly-defined loop at residues 82-86) is only 0.25 Å, approximately equal to the expected error calculated by Luzzati analysis of the zonal R-factor[13]. Only two sidechains appear to change conformation as a result of the mutation. Glu-109 swings down to interact with the positively-charged Arg-112 side chain while Arg-61, normally partially filling the space above Cys-112, moves out of the way to accommodate the new arginine side chain.

While wild-type apo-E favors binding to HDL over VLDL, the E4 mutation reverses this preference. This study of the N-terminal domain of apo-E4 does not indicate any significant structural changes that could account for altered lipoprotein binding, suggesting an indirect mechanism for this mutant's dysfunction. Studies with thrombolytic fragments of apo-E have shown that lipoprotein binding is mediated by the C-terminal domain[14] (despite the fact that the N-terminal domain is able to bind to artificial phospholipid disks[15]). One possibility is that the E4 mutation changes the interaction between the two domains, thereby altering the structure or accessibility of the C-terminal domain. Consistent with this interpretation, the lipid-binding properties of the 22-kDa fragment isolated from the E4 protein are indistinguishable from those of the wild-type fragment (unpublished observations). Several other studies have suggested communication between the two domains of apo-E, effectively linking the lipoprotein-binding and the receptor-binding functions[16]. The apo-E4 structure indicates that interdomain communication might be mediated in the absence of significant structural changes as have been observed for allosteric proteins.

TABLE 3.1. Statistics for data collection and structure refinement. Crystals of the E2 and E4 mutants were obtained via vapor diffusion by the hanging drop method. Conditions developed for the native apo-E3 protein[2] [15% PEG 400 (BDH), 20 mM sodium acetate, pH 5.3, 0.2% β-*n*-octylglucopyranoside (Calbiochem), and 0.1% β-mercaptoethanol] yielded suitable crystals for the E2 and E4 variants . All proteins crystallized in the $P2_12_12_1$ space group with approximately the same unit cell dimensions, a=40.7 Å , b=54.0 Å, c=85.4 Å. The *c*-dimension of the E2 mutant (83.91 Å) and the *b*-dimension of the E4 mutant (53.30 Å) are somewhat smaller than those of the wild-type protein. Data were collected using a RIGAKU automated four circle diffractometer (AFC5R) , equipped with an MSC cryo-cooling device. A 2.5 Å refined structure of the wild type 22-kDa fragment provided starting phases for the refinement of both mutants. The structures were refined using X-PLOR, version 2.0 (employing a combination of simulated annealing molecular dynamics, B-factor and positional refinement, and manual rebuilding of the structure).

	Native	E2	E4
Resolution (Å)	2.5	3.0	2.5
Diffraction data			
Number of unique reflections	6899	2750	6535
$\langle I \rangle / \langle \sigma_I \rangle$	43.5	32.1	26.3
X-PLOR refinement statistics			
Number of protein atoms (non-hydrogens)	1172	1167	1154
Number of water molecules	76	0	101
R_{cryst} (overall)	0.172	0.195	0.175
r.m.s. deviations from ideality			
bond lengths (Å)	0.017	0.016	0.015
bond angles (°)	3.2	3.7	3.1

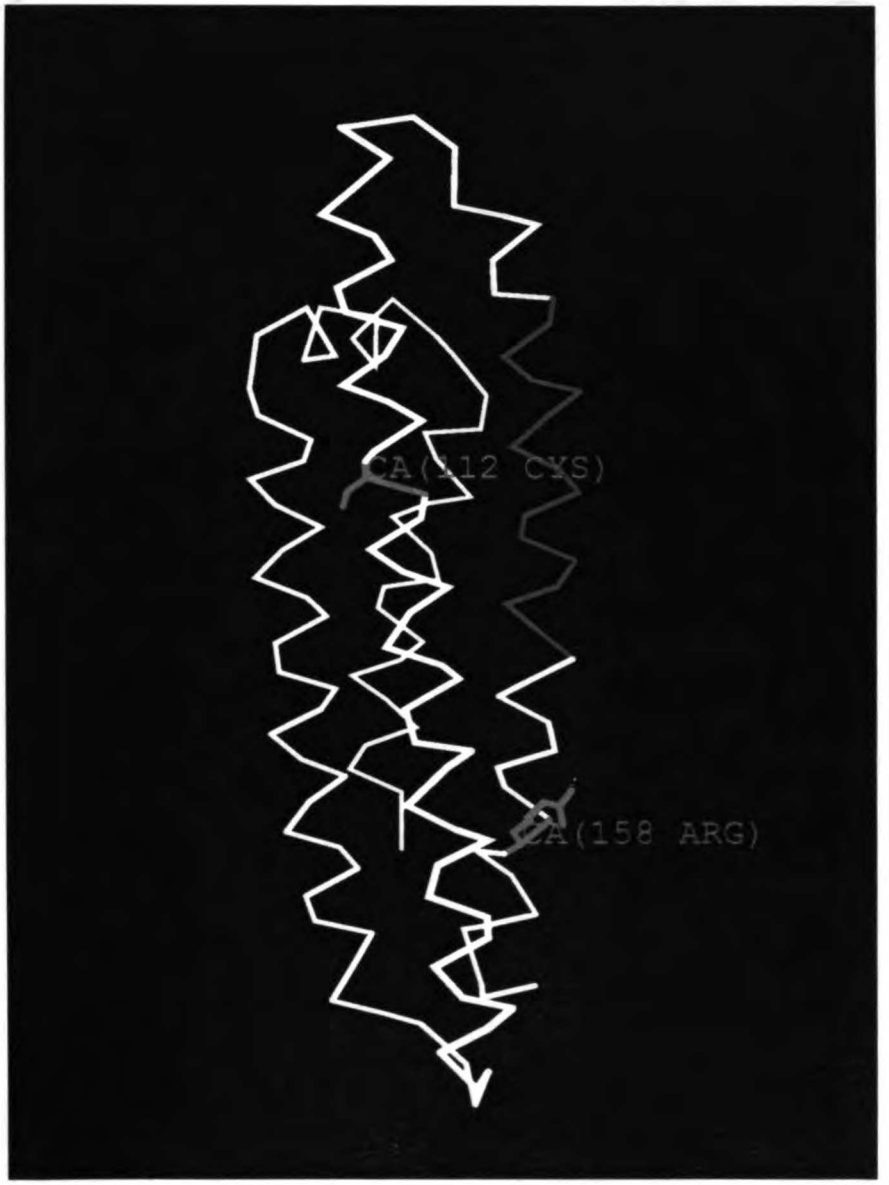


Figure 3.1. Ribbon diagram of the LDL receptor binding domain of apo-E. Residues 1-22 and 165-191 appear to be disordered and have not been modelled. Sites of the E4 (Cys-112) and E2 (Arg-158) mutations are indicated. Basic amino acids in the region 136-150 (shaded) are known to be important for LDL receptor binding.

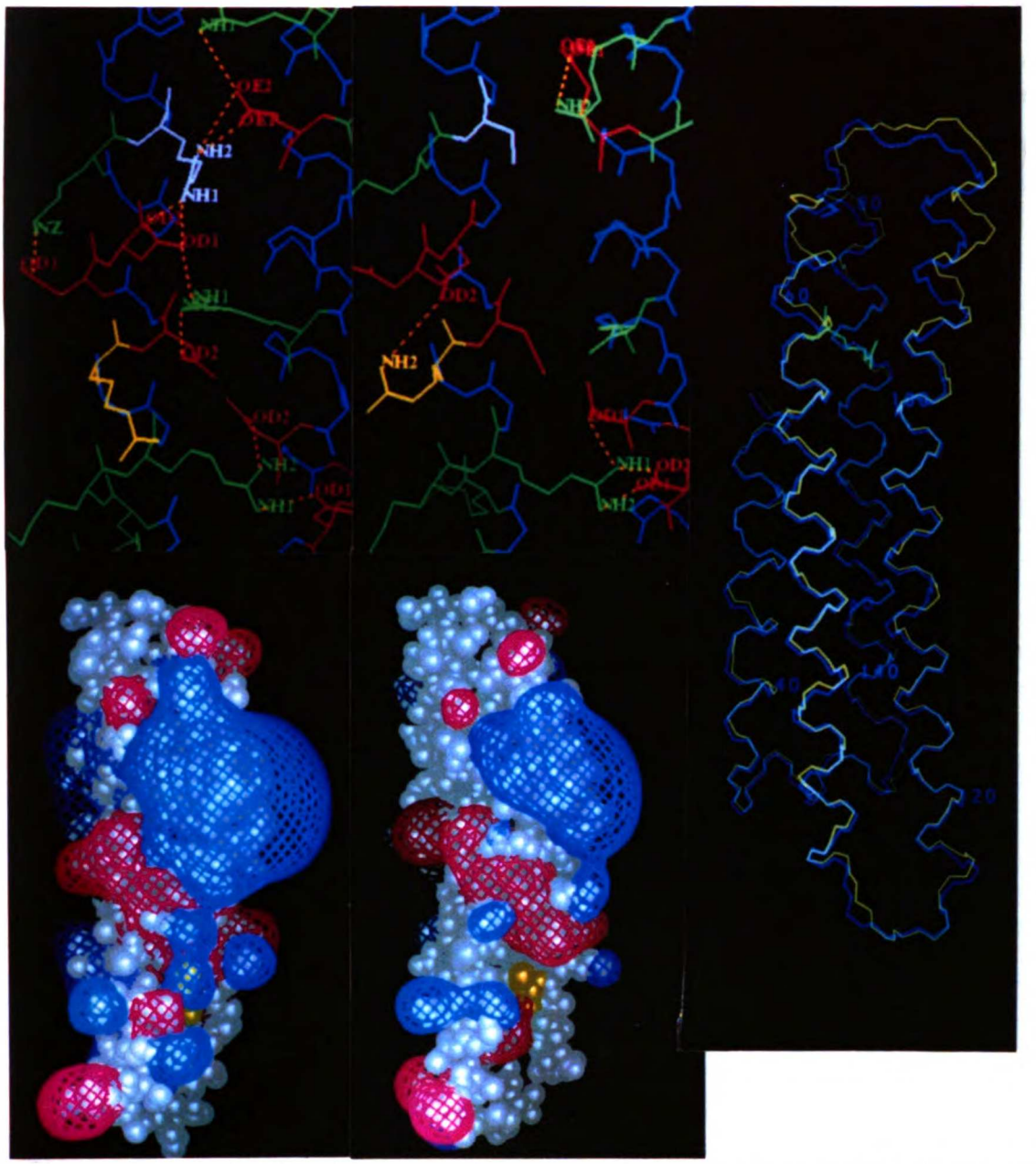


Figure 3.2. Structure of the apo-E2 mutant. The zone surrounding residue 158 is shown for the wild-type protein (a) and for the E2 mutant (Arg-158→Cys) (c). Salt bridges (defined as pairs of oppositely charged atoms separated by less than 5 Å) are indicated as stippled lines. Electrostatic potential map calculated for the wild-type (b) and apo-E2 mutant (d) proteins. The DELPHI program (Biosym, San Diego, CA) was used to calculate an approximate solution to the linearized Poisson-Boltzmann equation. The protein and solvent dielectrics were set to 2 and 80 respectively. The ionic strength was set to 150 mM (mimicking blood plasma). Only formal protein charges were included in the calculation. Positive (cyan) and negative (red) contours in the potential are evaluated at +2 and -2 kT per electron respectively. e) Superposition of mainchain atoms for the wild-type (blue) and apo-E2 (yellow) mutant proteins.

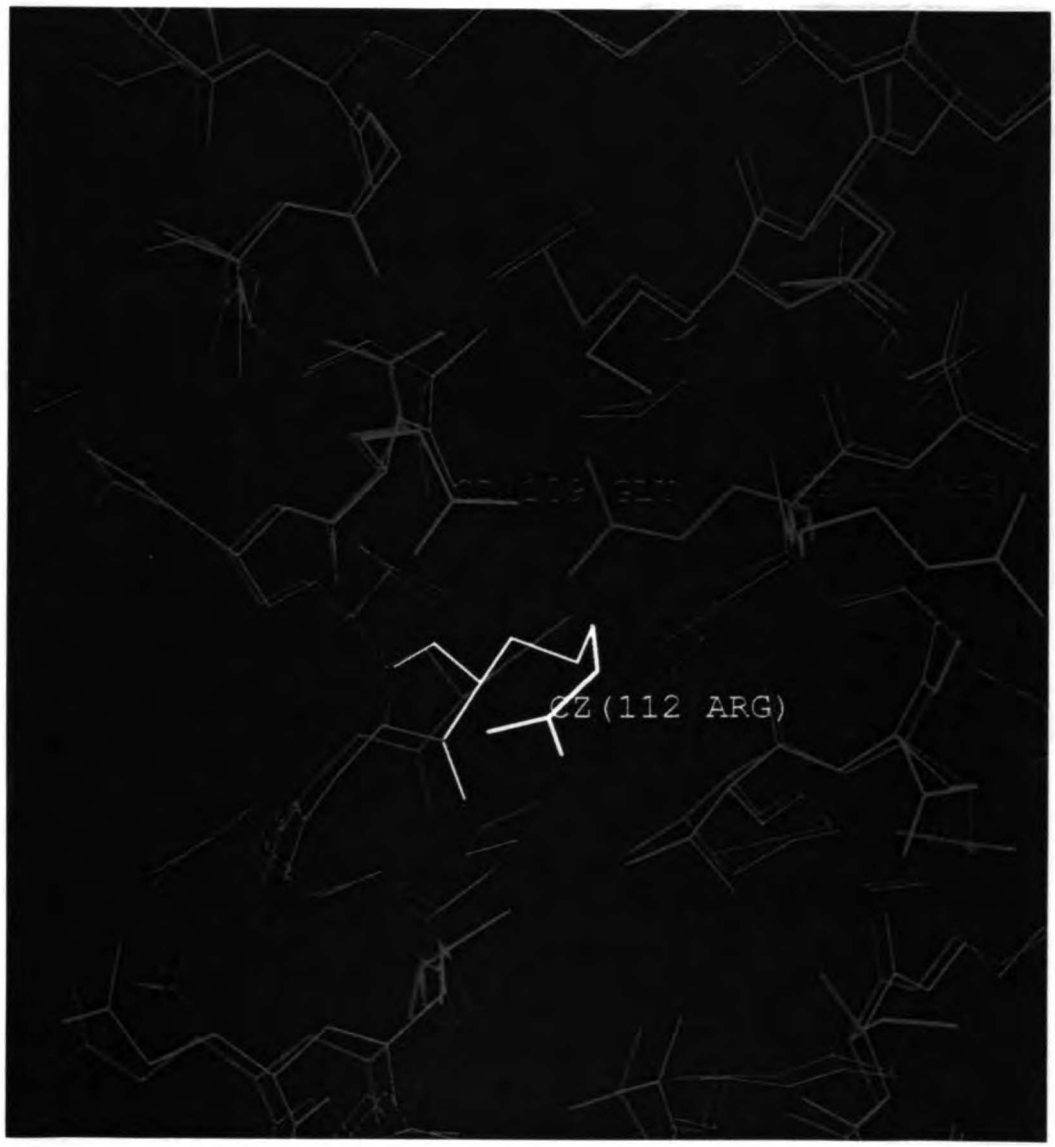


Figure 3.3. Structure of the apo-E4 mutant. Residues surrounding amino acid 112 are shown for the wild-type (magenta) and apo-E4 mutant proteins (green). Arg-112 in the E4 structure (white) forms a salt bridge with Glu-109, while Arg-61 (covering Cys-112 in the wild-type protein) swings out to accommodate the larger mutant side chain.

References

[1] Davignon, J., Gregg, R.E., & Sing, C.F. *Arteriosclerosis*, **8**, 1-21 (1988).

[2] Wilson, C., Wardell, M.R., Weisgraber, K.H., Mahley, R.W., & Agard, D.A. *Science*. in the press (1991).

[3] Mahley, R.W. *Science* **240**, 622-630 (1988).

[4] Uttermann, G., Langenbeck, U., Beisigel, U., & Weber, W. *Am. J. Hum. Genet.* **32**, 339 (1980).

[5] Rall, S.C., Jr., Weisgraber, K.H., Mahley, R.W. *J. Biol. Chem.* **257**, 4171 (1982).

[6] Weisgraber, K.H., Innerarity, T.L., & Mahley, R.W. *J. Biol. Chem.* **257**, 2518-2521 (1982).

[7] Steinmetz, A., Jakobs, C., Motzny, S., & Kaffarnik, H. *Arteriosclerosis* **9**, 405-411 (1989).

[8] Sing, C.F. & Davignon, J. *Am. J. Hum. Genet.* **37**,268-285 (1985).

[9] Miida, T. *Tohoku J. Exp. Med.* **160**, 177-187 (1990).

[10] Ehnholm, C., Lukka, M., Kuusi, T., Nikkila, E., & Utermann, G. *J. Lipid Res.* **27**, 227-235 (1986).

[11] Wetterau, J.R., Aggerbeck, L.P., Rall, S.C., & Weisgraber, K.H. *J. Biol. Chem.* **263**, 6240-6248 (1988).

[12] Aggerbeck, L.P., Wetterau, J.R., Weisgraber, K.H., Mahley, R.W, & Agard, D.A. *J. Mol. Biol.* **202**, 179-181 (1988).

[13] Luzzati, V. *Acta Cryst.* **5**, 802-810 (1952).

[14] Aggerbeck, L.P., Wetterau, J.R., Weisgraber, K.H., Wu, C-S.C., & Lindgren, F.T. *J. Biol. Chem.* **263**, 6249-6258 (1988).

[15] Innerarity, T.L., Friedlander, E.J, Rall, S.C., Weisgraber, K.H., & Mahley, R.W. *J. Biol. Chem.* **258**, 12341-12347 (1983).

[16] Wardell, M.R., Weisgraber, K.H., Havekes, L.M., & Rall, S.C. *J. Biol. Chem.* **264**, 21205-21210 (1989).

Chapter 4 :

Automated crystallographic phase refinement by iterative skeletonization

Abstract

A phase refinement procedure based on iterative skeletonization of electron density maps is presented. As with traditional solvent flattening methods, refinement alternates between real space and reciprocal space representations of the scattering density. A pseudo-atom list derived from the modified skeleton of an initial electron density map provides calculated structure factor amplitudes and phases. Recombination with the observed | F |'s yields a new map which can serve as the starting point for another round of skeletonization. Tests using partial structures to provide starting crystallographic phases has shown that this refinement procedure has a significantly larger radius of convergence than solvent flattening.

Introduction

Finding a solution to the crystallographic phase problem is generally the rate limiting step in macromolecular structure determination by x-ray diffraction (Blundell, 1976). All protein structures that have been solved crystallographically have relied upon a limited number of experimental or computational methods to obtain phase estimates prior to the building of a complete atomic model. These techniques, including isomorphous replacement, anomalous scattering, and molecular replacement, often yield poor initial estimates of the phases, resulting in electron density maps that are difficult to interpret. The development of techniques capable of improving phases in the absence of a complete atomic model should dramatically increase the rate of structure determination by x-ray crystallography and minimize problems such as incorrect chain-tracing that have plagued several recent structures (Branden, 1990).

The method of density modification (also known as solvent flattening) has been used in a number of cases to maximize the information available from a single derivative or from anomalous scattering data (examples include Fury, 1987; Tulinsky, 1988; Messerschmidt, 1989). Density modification is an iterative procedure consisting of two distinct steps (Wang, 1985). Initially, an electron density map (calculated using the current phase estimates) is filtered such that the density of all regions lying outside a molecular boundary is set to some mean value. Subsequent transformation of the filtered map yields a new set of crystallographic phases which can be recombined with the initial phases to produce an electron density map for the next cycle of refinement. While often successful, density modification has a limited radius of convergence, especially when the solvent content of a crystal is low (Zhang, 1990)[5]. In addition, since the procedure always converges to some solution, it is sometimes impossible to tell whether solvent flattening has improved the phases or not.

We have sought to develop new methods for phase refinement that can tolerate significantly larger errors in the starting phases. Our strategy has been to use the same cycle of real space—reciprocal space refinement, but to apply considerably stronger constraints on the real space representation of the scattering. These constraints have been applied by first skeletonizing the electron density map and then forcing this skeleton to adopt 'protein-like' characteristics. To evaluate our method, we have developed a number of test cases in which starting phases are derived from a partial atomic model. In the examples we have considered, our refinement procedure is able to converge to the correct phase solution, allowing the complete model to be built into the electron density map. In the same tests, solvent flattening produces a marginal improvement in the phases but leaves significant biases and errors in the final density map. Ultimately, we hope that this refinement scheme can be used to extend the application of molecular replacement techniques to proteins for which there is no close structural homolog, thereby obviating the requirement for experimental phase information.

Refinement method

Figure 4.1 outlines the key steps to our iterative phase refinement protocol, PRISM (Phase Refinement by Iterative Skeleton Modification). The procedure uses a set of experimental structure factors (F_{obs}) and corresponding initial phase estimates (α_0) to start the refinement. In the test cases we have considered, the α_0's are calculated from a partial model. Starting phases could, in principle, be derived from more conventional sources such as SIR, MIR, anomalous scattering, or molecular replacement using a more complete model. From F_{obs} and α_0, one can calculate a starting electron density map corresponding to the asymmetric unit in the unit cell. Each cycle of refinement uses an electron density map as input and ends with the synthesis of another map based upon (hopefully) improved

phases. By repeating several cycles of refinement, the phases gradually converge to final predicted values. The steps in each cycle of refinement are described below.

Step one: The electron density map (calculated on an ≈1Å rectangular grid) is skeletonized using the MKSKEL program, part of the GRINCH package (Williams, 1982; adapted for VMS by M. Carson, University of Alabama). This program outputs both a listing of nodes, corresponding to one-dimensional local maxima in the electron density map, and a listing of the connections between nodes. The nodes are assumed to represent pseudoatoms in the structure, with the node connections defining the bonds between them.

Step two: The skeleton output by MKSKEL is automatically modified using the CONNECT program (available upon request from CW) such that the scattering density resembles a single chain rather than a group of disconnected atoms or groups of atoms. This program initially identifies connected graphs of nodes and immediately prunes out the smallest graphs. Nodes with only a single bond to all other nodes (termed endpoints) are then identified and used to generate connections between the remaining isolated graphs. To force connections between graphs, the endpoints are allowed to diffuse through the unit cell subject to the cost function:

$$C(i,k) = -\sum_{j} \frac{1}{d_{ij}} + w_i \cdot \rho(k)$$

where $C(i,k)$ is evaluated for the *i-th* endpoint at the *k-th* pixel in the density map, d_{ij} is the distance between the *i-th* and *j-th* endpoints, w_i is an adjustable weighting parameter, and $\rho(k)$ is the electron density of the *k-th* pixel. The summation over *j*-endpoints includes all those which do not belong to the same graph as the *i-th* endpoint. Minimization of the cost function forces endpoints to form connections (decreasing d_{ij}) while remaining in regions of high electron density (keeping $-\rho(k)$ small). w_i is initially set such that the starting gradient is as close to zero as possible, thus balancing the two terms. For each endpoint,

the cost function at all adjacent pixels is evaluated and the endpoint is moved to the neighboring position with the lowest cost. If the current position has the lowest cost, w_i is dropped, thus removing the bias towards remaining in good electron density and favoring connection formation. Once a pair of endpoints meet, the two isolated graphs are grouped into a single larger graph and the pair of endpoints are removed from the minimization. Minimization is stopped after a fixed number of cycles. Small graphs which remain unconnected to others are removed from the skeleton. The nodes corresponding to the connections between graphs are added to the skeleton which is then output as a list of atoms (all with the same scattering properties).

Step three: The pseudo-atom list output by CONNECT is used to calculate structure factors (F_c) and phases (α_1) for the next cycle (assuming a carbon scattering curve for the pseudo-atoms). The F_c are scaled to the observed structure factors (using zonal scaling).

Step four: A new electron density map is calculated using the updated phases, α_1's, and modified Fourier coefficients, $F_w = 2wF_o\text{-}F_c$ (w = Sim's weight; Sim, 1960). Maps calculated using the F_w coefficients are ideal for this procedure since they include the true scattering vector at full scale while minimizing the systematic noise directed along α_1 (see Appendix). Alternative arguments for using these coefficients have been provided, and appropriate correction factors to account for experimental errors have been derived (Read, 1986).

To test the above procedure, we used the first fifty residues of myoglobin as the unknown structure. F_{obs} and α_{obs} were calculated using coordinates obtained from the Protein Data Bank entry 1MBD (Phillips, 1978). The space group and unit cell of apolipoprotein-E, another all-helical protein, was used for most calculations ($P2_12_12_1$, a=41.26, b=54.51, c=87.09, $\alpha=\beta=\gamma=90°$; ref. Wilson, 1991).

Several models were constructed to test different possible biases in the starting models. The first model contained only C_α atoms for all fifty residues. While this model has a small fraction of the total scattering density (50 atoms versus 417 in the full structure), the electron density is well distributed and there are no regions which are systematically under-represented. A second model included all backbone and C_β atoms for the first twenty-five residues. While this half-backbone test case includes a higher fraction of the scattering density (122 atoms), the distribution is highly asymmetric with the entire C-terminal half missing from the initial map. To test the effect of measurement errors and coordinate errors on refinement, both the observed structure factors (*i.e.* those calculated for the full 50 residue model) and the starting model coordinates were perturbed by the addition of a Gaussian distribution of random shifts.

The ability of the refinement scheme to converge to the correct electron density map was estimated by several criteria. Because the refinement is done using calculated "observed" data, the phase errors can be calculated exactly using the known structure. To estimate the accuracy of the phases, we have calculated the weighted phase error,

$$\Delta\Phi = \Sigma F_{obs} |\alpha_{obs} - \alpha_c| / \Sigma F_{obs}$$

where the summation is done over all reflections. The crystallographic R-factor between the scaled calculated F's and the experimental F's,

$$R = \Sigma | F_{obs} - F_c | / \Sigma | F_{obs} |$$

provides a useful measure of the error in the electron density map in cases in which the observed phases are not known.

Control experiments

The refinement procedure is based on the assumption that an electron density map can be accurately represented by the nodes in a skeleton. To confirm this assumption, the

following test was performed. An electron density map (≈1 Å grid spacing) was calculated for the first fifty residues of myoglobin, and subsequently skeletonized using the MKSKEL program (Williams, 1982). The nodes in the skeleton were treated as pseudo-atoms and used to obtain new structure factors and phases. Figure 4.2 shows the zonal R-factor between the correct F's (obtained from the Fourier transform of the true electron density map) and those calculated from the corresponding skeleton. The errors in the skeleton model are a strong function of the resolution of the initial map used to generate the skeleton. If the starting map is calculated directly from the atomic coordinates (*i.e.* with infinite resolution), the zonal R-factor remains below 25% for all data in the range 10-3 Å, but rises sharply at higher resolutions. A similar result is found using a 3 Å resolution starting map. If lower resolution starting maps are used to calculate the skeleton, the zonal R-factor remains at ≈25% for data below the resolution of the map but rises to that expected for a random structure at higher resolutions (Figure 4.2). If the correct map is used as the input to the PRISM procedure and refined for several cycles, the overall R-factor (10-3.0 Å data) rises to 30% while the phase error rises to 27°, providing a best case estimate for the final expected amplitude and phase errors.

Refinement tests with defined partial models

The first test of the refinement procedure used the C_α atom-only model of the myoglobin fragment as a starting partial model. An initial map calculated at 3 Å resolution using the observed structure factors and C_α-based phases is shown in Figure 4.3a. As described in METHODS, this map was skeletonized using the MKSKEL program (Williams, 1982; Figure 4.3b), the resulting skeleton was then modified by the CONNECT program (Figure 4.3c), and a new map was calculated with the modified skeleton providing new phases and modified structure factors (Figure 4.3c). By iterating this procedure ten

times, the final refined map shown in Figure 4.3d was produced. Figure 4.4 shows the change in overall R-factor and phase error during refinement. Whereas the initial phase error for the C_α model is 58°, refinement lowers this to 33.5°. Direct comparison of the true electron density map with the final map indicates that refinement restores the majority of the missing side chain density (allowing the side chains to be easily modelled), as well as removing the extraneous density lying outside the protein region. The R-factor and phase error in the final cycle closely approaches that obtained when refining from the true electron density map. Whereas the directionality of the starting model is ambiguous, density from the peptide carbonyl oxygens becomes readily apparent in the refined map, permitting the full atom backbone to be modelled.

To allow a direct comparison between PRISM refinement and conventional density modification, we have used the B.C. Wang solvent flattening programs (Wang, 1985) with the backbone-atom model providing a starting phase probability distribution for each reflection. The solvent fraction was set to 85% and ten cycles of iterative map filtering and phase recombination were carried out (recalculating the molecular boundary every cycle). The final recombined phases calculated by solvent flattening have an average phase error of 49.0° (Figure 4.4), only marginally better than the starting phase distribution (phase error = 58.0°). In this test case, at least, PRISM refinement is considerably more effective than the method of solvent flattening.

A more realistic test of the refinement procedure was constructed using a model containing only backbone atoms for the first twenty-five residues. Because this model lacks the majority of the side chain atoms and the entire C-terminal half of the molecule, it more accurately simulates a potential starting map that one might obtain from molecular replacement with a fragment search model. Figure 4.5 shows the starting and final maps for this test case. Similar improvements in the overall R-factor and the phase error are seen with this half-backbone model (Figure 4.6) as with the C_α-only model. The improvement

is especially dramatic in the C-terminal region (Figure 4.5c,d), with the disordered, unconnected density becoming continuous and easily interpretable in the final map. The final phase error in this test (27.2°) is essentially identical to that obtained after several cycles of refinement with the complete model as a starting structure (26.9°). In contrast, the solution obtained following extensive solvent flattening had an average phase error of 52.6° (Figure 4.6). Traditional density modification thus only slightly improves the phases relative to their initial estimates (phase error = 56.6°), and yields a final map that is significantly noisier than the PRISM-refined map, especially in the C-terminal region.

Understanding the requirements for phase convergence

Several additional tests were performed to simulate potential problems that might be encountered in a true structure determination. To approximate data collection errors, the "observed" structure factors were altered by the addition of a Gaussian distribution of random shifts such that the R-factor between the true data and the modified data was ≈7%. This level of noise, significantly more than one might expect with typical diffraction data, appeared to have only minor effects on the refinement procedure (slowing it slightly) when using the half backbone starting model (Table 4.1). Similar effects were observed when using a half-backbone starting model which had been perturbed by the application of 0.5 Å random shifts to the atomic coordinates (Table 4.1). While systematic errors are likely to cause artifacts in the maps, the refinement procedure does not appear to require data free of random errors or a perfectly accurate partial starting model.

To understand what aspects of the procedure are required for the phases to converge to the true values, we repeated the half-backbone test several times altering steps in the refinement protocol. If conventional $2|F_o| - |F_c|$ Fourier coefficients are used in place of the optimally weighted coefficients ($2w|F_o| - |F_c|$), the phases improve only

marginally after several cycles (Table 4.1). Modification of the skeleton also appears to play a key role; if the skeleton is not pruned or if connections in the density are not generated, the phases improve somewhat but the drop in the phase error is not as significant (Figure 4.7). The resolution of the starting electron density map is important for proper refinement; a map with less than 3.5 Å resolution does not converge whereas one calculated at higher than 3.5-Å converges easily (Table 4.1).

Discussion

This work describes the development of an automated phase refinement procedure which we envisage using in combination with traditional molecular replacement to solve the crystallographic phase problem in the absence of experimental information. To apply this algorithm to *de novo* protein structure determination, an initial guess for part of the structure must first be made. After using molecular replacement techniques to position this predicted structure in the crystallographic asymmetric unit, the algorithm is designed to refine an electron density map such that it converges to the true scattering density. Once the density map has been refined, a full atomic model can be built for subsequent refinement by conventional techniques. The key step in the successful application of the method is making a correct guess for the starting partial model. This step could in principle be done by a combination of protein structure prediction methods and interpretation of the experimentally-determined Patterson map. A Patterson map, yielding information about the interatomic vectors in a crystal structure, can be calculated from observed reflection intensities without additional phase information. By interpreting features in the Patterson map, one may obtain clues about the amount of secondary structure, the relative orientation of alpha helices and beta strands, and the spatial relationship between secondary structure elements (Blundell, 1976). Our method should allow the extension of traditional molecular

replacement techniques to proteins which do not share homology with any previously determined proteins.

As with standard density modification (solvent-flattening) techniques, our method alternates between real space and reciprocal space representations of the scattering density. Whereas solvent flattening uses the real space constraint that all density lie within a molecular boundary, our constraint is significantly more restrictive. By skeletonizing the map and using the nodes in the skeleton as dummy atoms, we immediately enforce positivity and atomicity on the electron density. The scattering density in a protein crystal can generally be represented as a single connected chain (Greer, 1985). By modifying the skeleton such that distinct groups of atoms are connected to one another, we force this constraint to be satisfied at every cycle in the refinement procedure. Because both the skeletonization and skeleton modification steps are implemented as computer programs, these relatively strong constraints on the electron density can be applied with no undue human bias.

The radius of convergence of any refinement procedure is likely to be a function of the ratio between the number of variables used to represent the scattering density and the number of structure factor observations. As this ratio increases, it becomes easier to fall into incorrect minima in the refinement cost function since errors in the electron density can be accurately modelled by the excess parameters. The skeleton representation requires relatively few parameters (the grid coordinates for each of the pseudo-atoms, approximately equal to the number of true atoms in the structure) but models the scattering density quite accurately (as shown in Figure 4.2). In contrast, traditional density modification techniques treat each electron density map pixel within the molecular boundary as a continuous variable during refinement. Using a 1-Å grid (appropriate for 3-Å data), a single atom contributes significant electron density to ≈30 surrounding pixels. The skeletonization procedure thus reduces the effective number of parameters by

approximately an order of magnitude. This effect alone may help explain the larger apparent radius of convergence of the PRISM method.

In addition to differences in the real space constraints, application of the reciprocal space constraints also differs between the solvent flattening and PRISM procedures. In Wang's implementation of solvent flattening (Wang, 1985), the filtered map phase probability distribution is multiplied at each cycle by the starting (experimentally-determined) probability distribution to arrive at a new estimate. In contrast, our procedure ignores all previous estimates and uses only the current electron density map to obtain phases. Reciprocal space recombination is achieved using modified Fourier coefficients ($2w\,|F_o|-|F_c|$), yielding a new map that has the full component of the missing density yet the minimal amount of systematic noise (*i.e.* errors directed along the α_{calc} vector). If experimental phase information exists, it should be possible to implement a reciprocal space recombination that uses this in each refinement cycle. The current procedure is appropriate, however, when the starting phase estimates are based on an incomplete atomic model which contains significant errors.

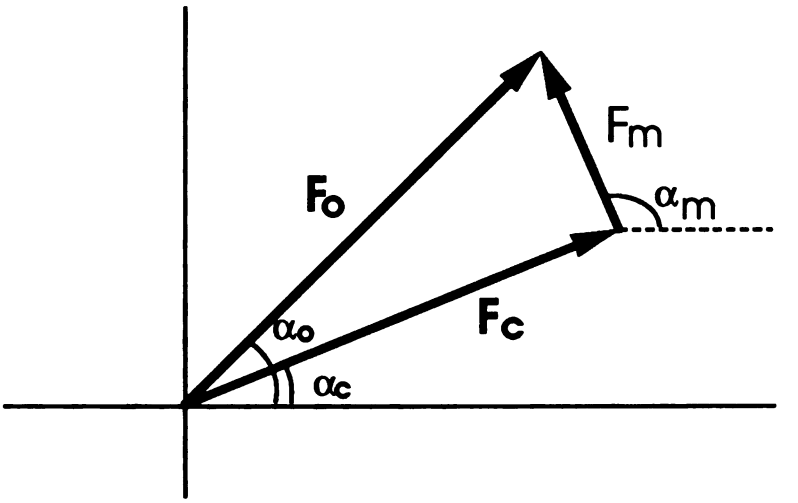
Future work using this methodology will incorporate additional constraints in the refinement procedure. For instance, PRISM refinement could be applied to only the poorly defined regions in a map derived from a relatively complete molecular replacement model. At each cycle, new phases would be obtained using the combined real atoms from the molecular replacement model and the non-overlapping pseudo-atoms from the CONNECT output. Alternatively, if a heavy atom derivative has been found, its phase information could be used in a standard recombination scheme to direct the refinement. The addition of constraints on the phases should extend the radius of convergence of the method and allow the use of even cruder partial models to start refinement.

Acknowledgements:

We wish to thank Dr. Lynn Ten Eyck for a conversation that inspired us to take this approach. Funding was provided by the Howard Hughes Medical Institute. CW was supported by a Fannie and John Hertz Foundation fellowship in applied physics.

Appendix: Optimal weighting scheme for modified Fourier coefficients

Given phases calculated for a partial structure, our goal is to find optimal Fourier coefficients that properly weight the contribution of the scattering from the missing atoms.



Model scattering $= F_c \exp i\alpha_c$

Missing scattering $= F_m \exp i\alpha_m$

True scattering $= F_o \exp i\alpha_o \quad = F_c \exp i\alpha_c + F_m \exp i\alpha_m$

A standard difference map (using $|F_o|-|F_c|$ coefficients and α_c phases) can be decomposed into three components, corresponding to the true missing signal, to systematic noise (directed along α_c), and to random noise (Blundell and Johnson, 1976):

$$(F_o-F_c) \exp i\alpha_c = [F_c/(F_o+F_c)]\ F_m \exp i\alpha_m \quad \textit{(scaled true signal)}$$

$$+ \quad [F_m^2/(F_o+F_c)] \exp i\alpha_c \quad \textit{(systematic noise)}$$

$$+ \quad \{F_cF_m/(F_o+F_c)] \exp i(-\alpha_m + 2\alpha_c) \quad \textit{(random noise)}$$

An optimal map can be obtained by subtracting the best estimate for the systematic noise ($[<F_m^2>/(F_o+F_c)]$) and by scaling the (Fo-Fc) map such that the true scattering is given its full weight ($1 / [F_c/(F_o+F_c)]$). To estimate the average systematic noise (Sim, 1960):

$$\frac{<F_m^2>}{(F_o+F_c)} = \frac{F_o^2 + F_c^2 - 2F_oF_c \int p(x)\cos x\, dx}{(F_o+F_c)} = \frac{F_o^2 + F_c^2 - 2F_oF_cw}{(F_o+F_c)}$$

where $w = I_1(X) / I_0(X)$, $X = 2F_oF_c/<\Delta F^2>$. After subtracting the average noise, scaling the true signal, and adding the model scattering (to give the complete structure rather than the difference signal):

$$[(F_o-F_c) \exp i\alpha_c - <F_m^2>/(F_o+F_c) \exp i\alpha_c] / [F_c/(F_o+F_c)] + F_c \exp i\alpha_c =$$
$$[(F_o-F_c)(F_o+F_c) - F_o^2 - F_c^2 + 2F_oF_cw + F_c^2] / F_c \exp i\alpha_c =$$
$$[2wF_o - F_c] \exp i\alpha_c$$

Thus, a map calculated with weighted coefficients, $F_w = 2w|F_o| - |F_c|$, will provide the full scattering signal with the correct phase, plus additional random and systematic noise which has been minimized given that the true phase is unknown.

$$(2wF_o - F_c) \exp i\alpha_c = F_m \exp i\alpha_m + F_c \exp i\alpha_c \quad \textit{(true signal at full strength)}$$
$$+ (F_m^2 - <F_m^2>)/F_c \exp i\alpha_c \quad \textit{(minimal systematic noise)}$$
$$+ F_m \exp i(-\alpha_m + 2\alpha_c) \quad \textit{(random noise)}$$

Table 4.1: Requirements for phase refinement.

Conditions of the test	Starting R-factor	Starting phase error	Final R-factor	Final phase error
Standard refinement	50.5%	60.1°	30.5%	27.2°
Using 2Fo-Fc recombination	"	"	37.0%	42.4°
Not pruning small groups of atoms	"	"	33.9%	36.3°
Not generating connections	"	"	34.5%	40.4°
Addition of 7% noise to the data	52.4%*	"	34.3%*	33.1°
0.5Å random shifts of starting model	56.35%	61.9°	33.2%	33.4°
Low resolution (4Å) starting map	50.5%†	61.6°†	39.7%	53.4°

The half-backbone atom test (using residues 1-25 backbone + C_β atoms to model the first 50 residues of myoglobin) was carried out using the standard refinement scheme described in the test. In addition, the refinement was repeated several times with slight changes in the refinement protocol as described above. *- R-factor calculated with F_o = error-added data, not error-free data. †- starting parameters calculated after the first cycle of skeletonization.

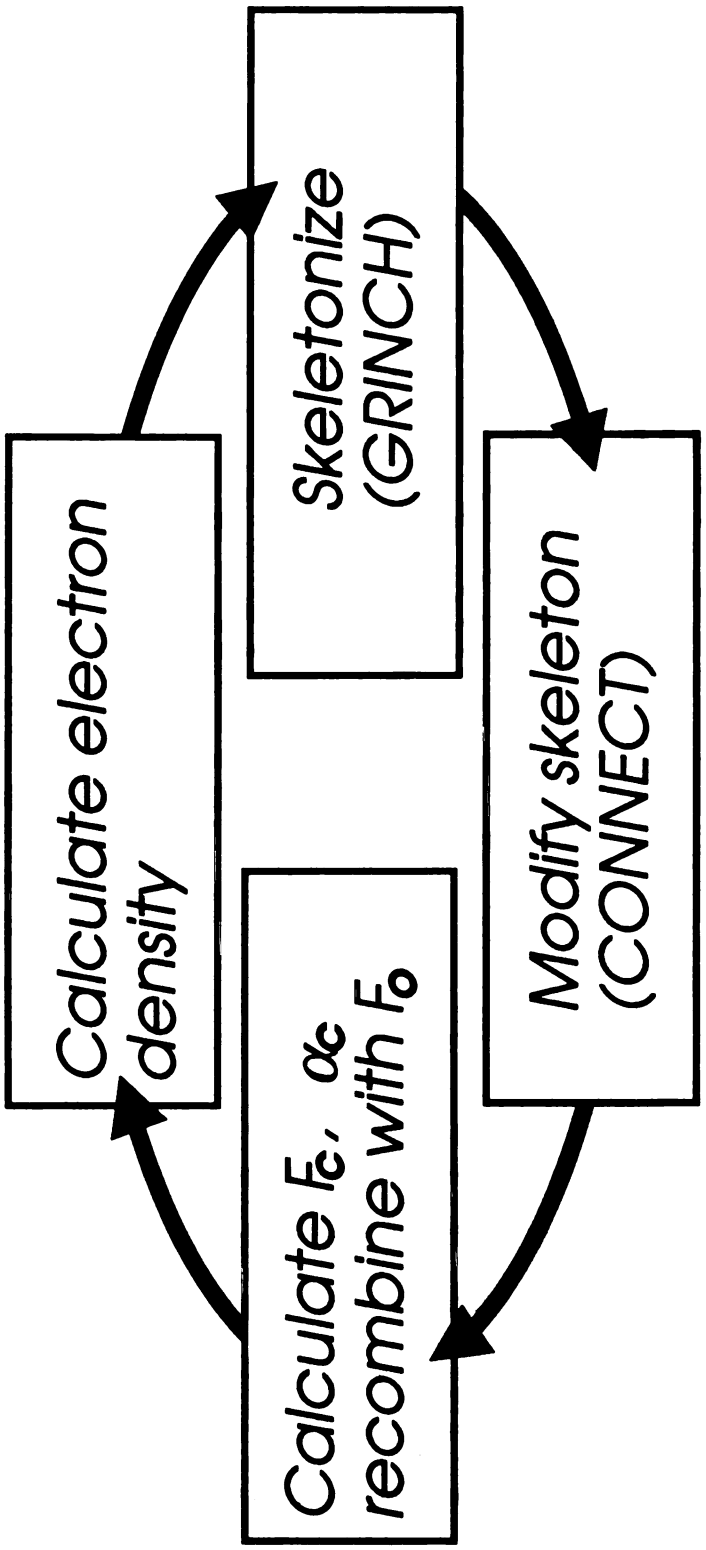


FIGURE 4.1. Overview of the PRISM phase refinement scheme. Details for each step are described in METHODS.

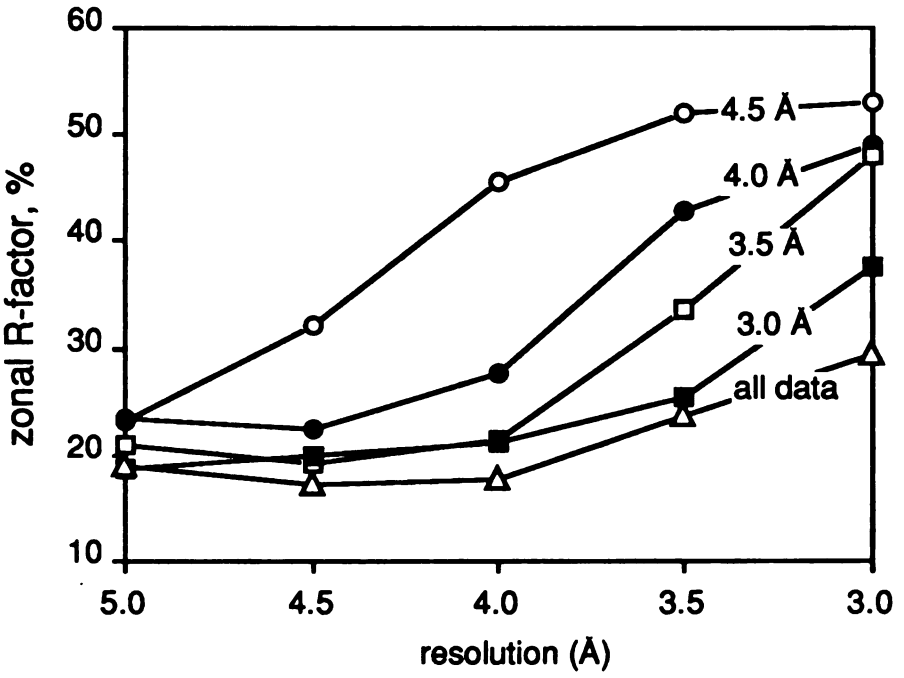


FIGURE 4.2. Accuracy of the skeletonization procedure. Structure factors calculated from the atomic model for the first fifty residues of myoglobin were used to calculate a series of electron density maps with decreasing resolution. The maps were subsequently skeletonized using the GRINCH MKSKEL program. The unfiltered output (a list of local maxima in the density map) was used to calculate new structure factors (each local maximium was assumed to represent a dummy carbon atom). Figure shows the zonal R-factor as a function of zonal resolution for each of the maps.

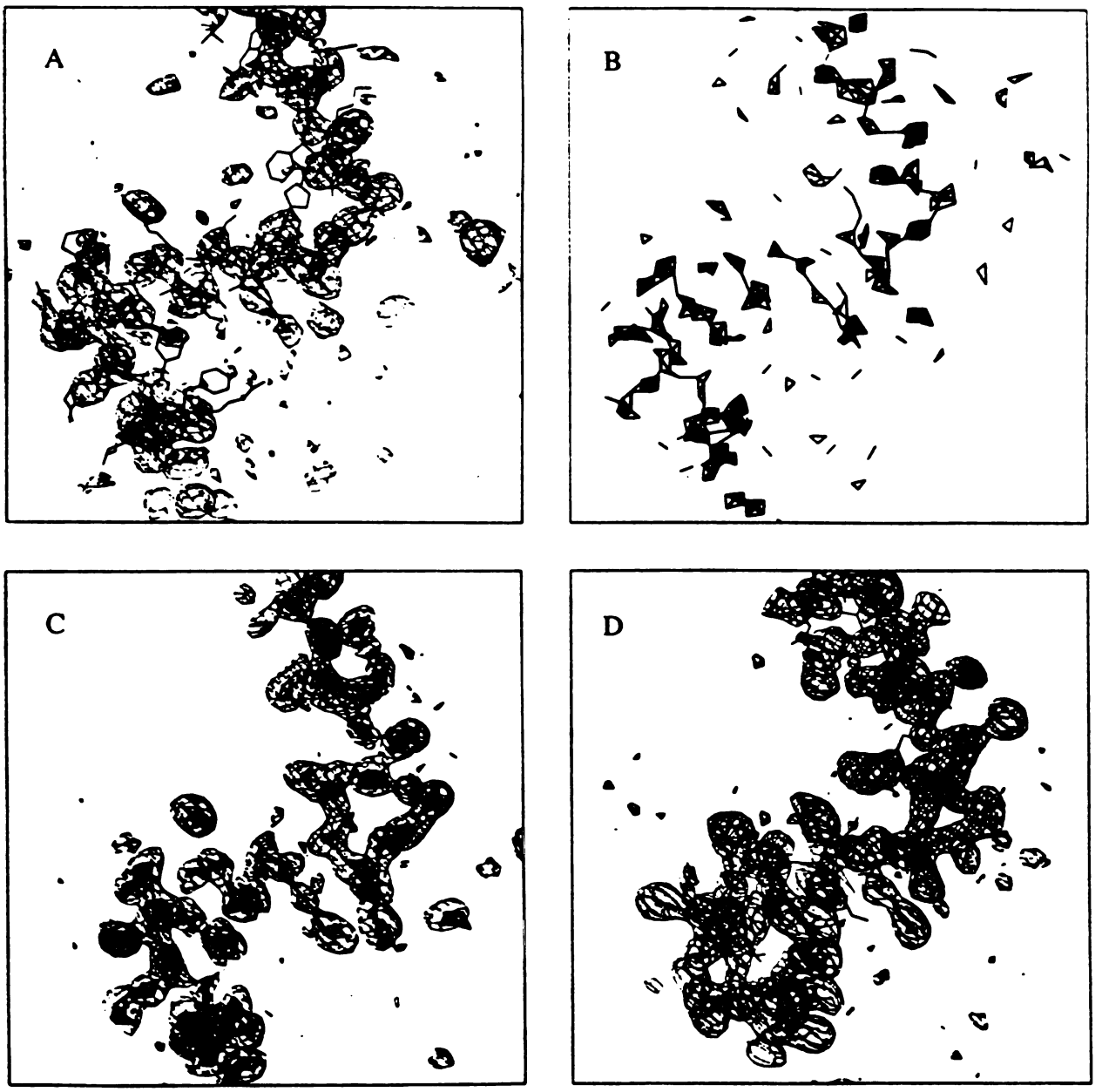


FIGURE 4.3. Steps in the refinement of a C_α-atom-only starting model. a) Starting electron density map calculated using phases derived from the C_α-atom model. b) Skeleton produced from the map in (a) using the GRINCH MKSKEL program. c) Skeleton following modification by the CONNECT program. Electron density is calculated using phases derived from the modified skeleton. d) Electron density calculated after ten cycles of PRISM refinement.

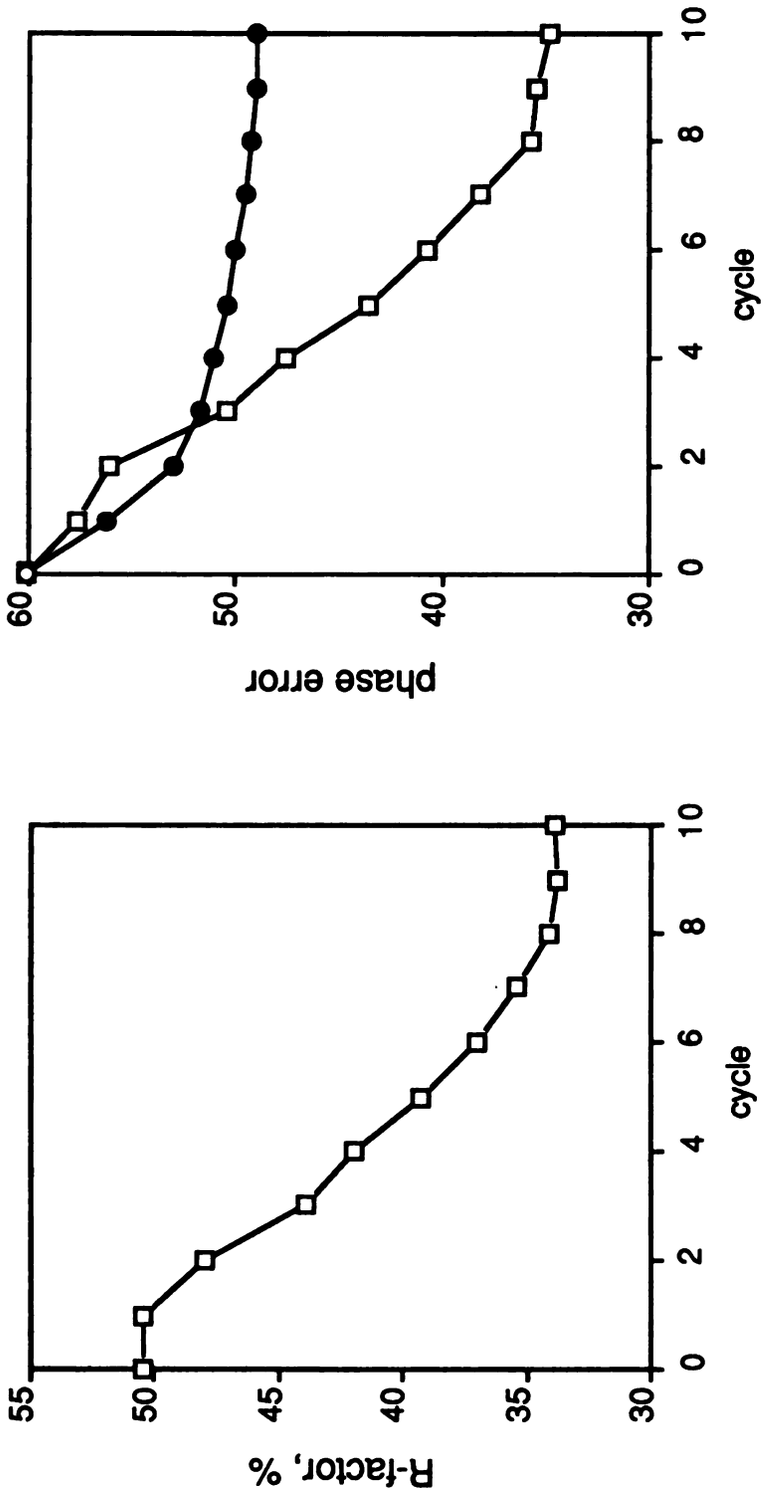


FIGURE 4.4. Results for C_α-atom-only starting model. Overall R-factor and phase error calculated as a function of refinement cycle, using the C_α-atom model to start refinement of the myoglobin fragment. Phase error is shown for both the PRISM refinement (open squares) and the solvent-flattening refinement (solid circles).

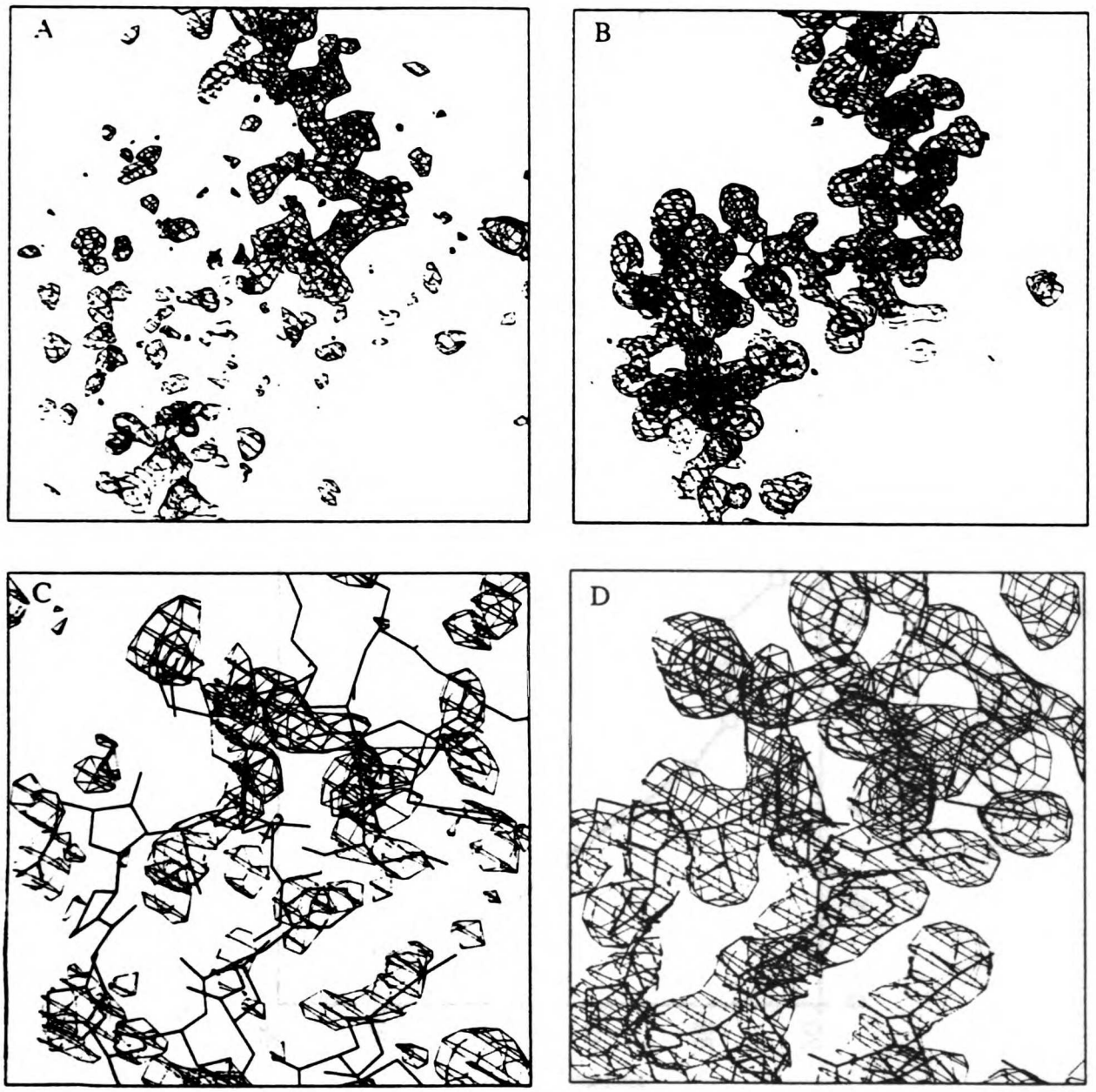


FIGURE 4.5. Starting and final maps using a half-backbone-atom starting model. a) Starting electron density map calculated using phases derived from the half-backbone-atom model. b) Final electron density map after ten cycles of PRISM refinement. c,d) C-terminal region of the electron density before (c) and after (d) PRISM refinement.

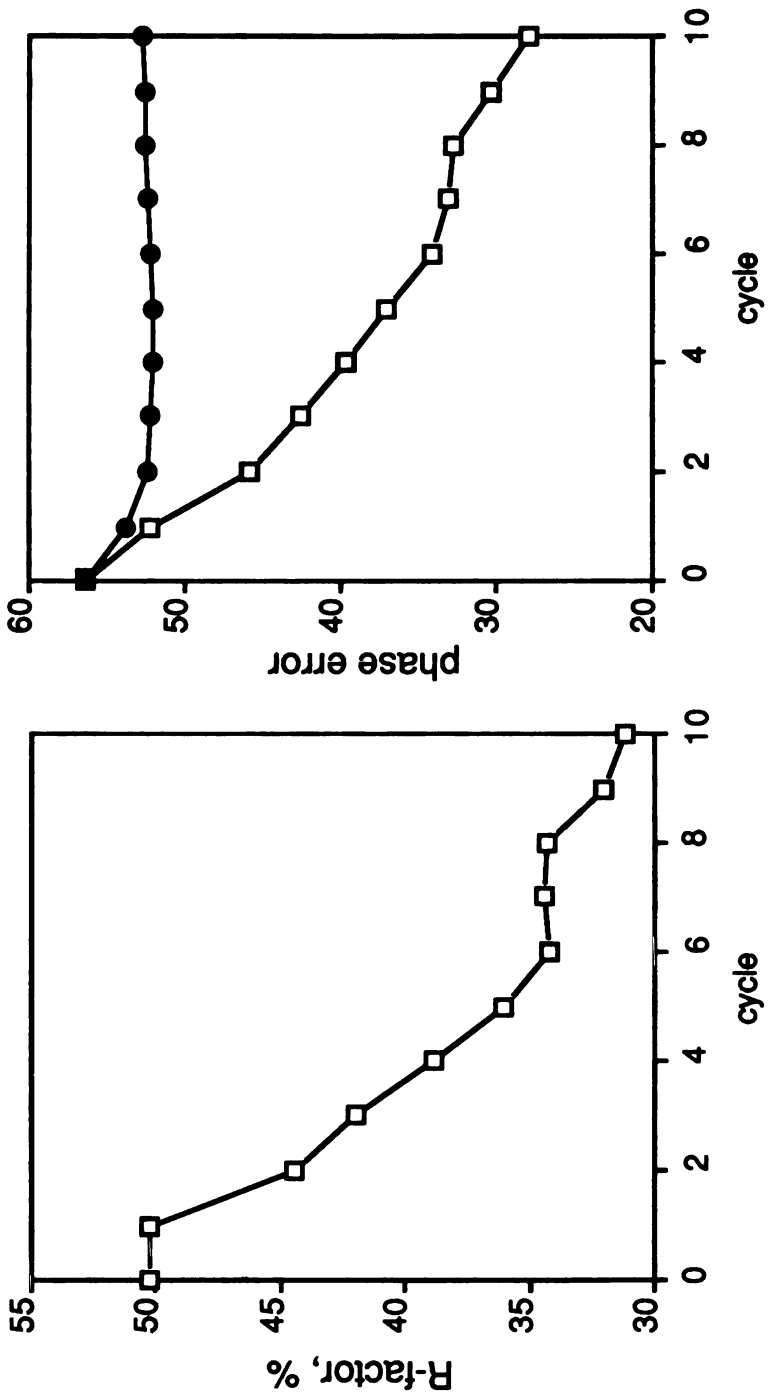


FIGURE 4.6. Results for half-backbone-atom starting model. Overall R-factor and phase error as a function of refinement cycle, using the half-backbone atom model to refine the myoglobin fragment. Phase error is shown for both the PRISM refinement (open squares) and the solvent-flattening refinement (solid circles).

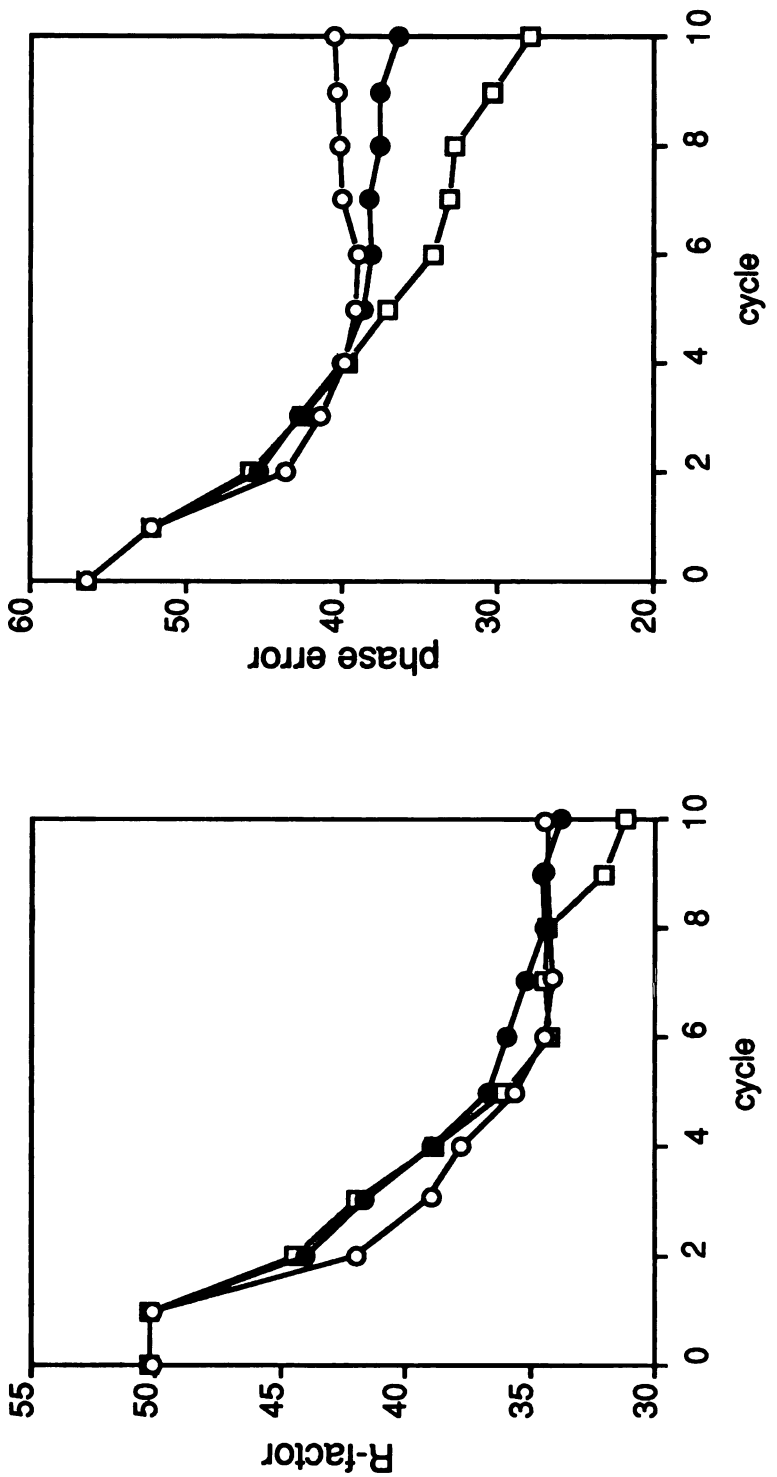


FIGURE 4.7. Effects of skeleton modification on refinement. The half backbone atom model was used to start refinement of the myoglobin fragment. In addition to the standard protocol (open squares), the refinement was also carried out in tests in which 1) connections between isolated subgraphs were not generated (open circles), or 2) small graphs were not pruned out of the skeleton (closed circles).

References

Blundell, T.L. & Johnson, L.N. (1976) Protein Crystallography. Academic Press. New York.

Branden, C.-I. & Jones, T.A. (1990) *Nature.* **343**, 687-689.

Fury, W.F., Robbins, A.H., Clancy, L.L., Winge, D.R., Wang, B.C., & Stout, C.D. (1987). *Experientia Supplementum* **52**, 139.

Greer, J. (1985) *Methods in Enzymology.* **115**, 206.

Luzzati, V. (1953) *Acta Cryst.* **6**, 142.

Main, P. (1979) *Acta Cryst.* **A35**, 779.

Messerschimdt, A., Rossi, A., Ladenstein, R., Huber, R., Bolognesi, M., Gatti, G., Marchesini, A., Petruzzelli, R., & Finazzi-Agro, A. (1989). *J. Mol. Biol.* **206**, 513.

Phillips, S.E.V. (1978) *Nature.* **273**, 247.

Read, R.J. (1986) *Acta Cryst.* **A42**, 140.

Sim, G.A. (1960) *Acta Cryst.* **A13**, 511.

Tulinsky, A., Park, C.H., & Skrzypczak-Jankun, E. (1988). *J. Mol. Biol.* **202**, 885.

Wang, B.-C. (1985) *Methods in Enzymology.* **115**, 90.

Williams, T. (1982). *A man-machine interface for interpreting electron density maps.* (Ph.D. thesis, University of North Carolina, Chapel Hill).

Wilson, C., Wardell, M., Weisgraber, K.H., Mahley, R.W., & Agard, D.A. (1991). *Science.* in press.

Chapter 5 :

A computer model to dynamically simulate protein folding: Studies with crambin

Abstract

The current work describes a simplified representation of protein structure with uses in the simulation of protein folding. The model assumes that a protein can be represented by a freely-rotating rigid chain with a single atom approximating the effect of each sidechain. Potentials describing the attraction or repulsion between different types of amino acids are determined directly from the distribution of amino acids in the database of known protein structures. The optimization technique of simulated annealing has been used to dynamically sample the conformations available to this simple model, allowing the protein to evolve from an extended, random coil into a compact globular structure. Many characteristics expected of true proteins, such as the sequence-dependent formation of secondary structure, the partitioning of hydrophobic residues, and specific disulfide pairing, are reproduced by the simulation, suggesting the model may accurately simulate the folding process.

Introduction

Despite years of both experimental and theoretical study, predicting the three-dimensional structure of a protein from its primary sequence is still a major unsolved problem. The need for a solution has increased in recent years with advances in DNA sequencing technology: the rate at which protein sequences are being determined by molecular biologists far exceeds the rate at which crystallographers can solve protein structures. While empirical force fields have been developed which can accurately predict the conformation of small molecules or subtle changes in folded protein structures (via molecular mechanics or molecular dynamics), these descriptions have not been useful for the simulation of protein folding because their highly-dimensioned potential surfaces are characterized by many local minima. Solving the protein folding problem with these force fields would require the global optimization of thousands of variables which are non-linearly coupled to each other. Early computer studies by Levitt and Warshell[1], Scheraga *et al* [2], and Kuntz *et al* [3] showed the advantages of using simplified protein models. By limiting the number of structural degrees of freedom to one or two per amino acid, the conformation space which must be sampled to arrive at the final folded structure is significantly smaller than that available to a full-atom model of the protein.

Over the last ten years, little new work has been done to explore the potential power of simplified models like those described in these initial studies. There are several reasons why this basic approach should be reconsidered. 1) The size of the protein structure database which is used to develop and test these models has increased significantly. 2) Several new approaches to multivariate optimization problems, such as simulated annealing and neural networks, have been developed which may solve the local minimum problem for simplified models. 3) The speed of computers available for these studies has increased by several orders of magnitude, allowing a more complete search of conformation space

(their speed would still have to improve by a factor of $>10^{10}$ if molecular dynamics on a full-atom model were used to fold a protein!). Historically, these models fell out of favor after it was shown that unreasonable, built-in biases in the early models allowed them to work for specific test proteins[4]. Hagler and Honig demonstrated that an extremely simplified model for bovine pancreatic trypsin inhibitor (BPTI), in which the protein consisted of only glycine and alanine residues, produced a folded structure which was as correct as that generated by the Levitt-Warshel model. In addition, they showed that many topologically important features of BPTI, such as the 180° twist in its beta sheet, were not reproduced by any of the computer folding attempts. They conclude that by assuming certain properties about the starting conformation and manipulating terms in the force field used to evaluate different conformations, it is not hard to produce folded structures that bear some resemblence to a test protein. An important unanswered question, however, is whether a well-conceived, unbiased simplified model can contain enough structural and energetic information to truly predict protein structure.

The current work describes a simple representation for the structure and energetics of a protein. We have tried to avoid the potential problems of unwarrented bias in the model by relying completely on the statistical preferences observed in the database of known crystal structures to derive all aspects of the model. Completely random starting structures are minimized using simulated annealing. In no way is the crystal structure of our test protein (crambin) used to define the model. Hopefully, therefore, our findings can be generalized to a large extent to other proteins. Several results indicate that the model is able to reproduce the characteristics of a folded protein in a sequence-dependent way.

Crambin was chosen as a test protein because its crystal structure is known to high resolution and it is one of the smallest proteins with significant amounts of secondary structure. Different aspects of the correct crambin structure can be analyzed and compared to the results of the folding simulation. The agreement or disagreement between the two

sets of data can then be used to assess the quality of the model. In a number of ways, listed below, the folding simulation is able to reproduce the features of the true crambin structure. 1) Alpha-helices and beta strands develop where they occur naturally in the crystal structure. 2) The structures produced by the minimization are reasonably similar to the true structure in terms of their r.m.s. deviation or distance matrix error (to our knowlege, their average deviation is the lowest ever obtained by the simulated folding of a protein). 3) The formation of certain disulfides, corresponding to the true disulfides, is significantly favored over non-native disulfides. The simulation with the true crambin amino acid sequence and a variety of control experiments using related sequences suggest that these results are not dependent on fortuitous biases within the model. For instance, completely different structures result when the crambin sequence is randomized, maintaining the same amino acid composition. The tendency for secondary structure to form is largely lost and that which does form is quite different when folding individual segments of the sequence, as might have been expected from studies with small peptides. Starting from different random models, simulated annealing produces a collection of closely related folded structures, indicating the ability of the method to avoid many local minima during the journey down the potential surface towards the global minimum. Several properties of the native folded models, such as the tendency for hydrophobic residues to partition into the interior of the protein, indicate that the folding process produces 'protein-like' structures. Taken together, these results strongly encourage the use of simplified models for protein structure, both for the purpose of protein structure prediction and as a means of studying the energetics and dynamics of protein folding.

While this model certainly does not solve the protein folding problem, it suggests that a large part of the physics which is important for defining protein structure can be modelled by united atom residues and simple pairwise potentials.

Methods

Several models have been used in the simulation of protein folding. Lattice models, in which each residue is constrained to lie at a lattice point with appropriate connections to adjacent residues, have been used because it is computationally possible to generate a significant fraction of the allowed conformations. The energy of a given conformation for a lattice model is generally calculated as the sum of interactions between adjacent or near-neighbor lattice sites, the value of each interaction determined by the type of interacting residues. Go and Taketomi represented space as a two-dimensional lattice and were able to hypothesize a role for short- and long-range forces in the folding transition[5]. The highly-simplified geometry of such models, however, lacks the basic structural properties of true proteins and thus compromises the accuracy of the model for the simplicity with which conformations may be generated.

The models used for simulations in the present work have attempted to maintain a minimum of structure while preserving the basic geometry of the polymer. All backbone bond lengths and bond angles were held at ideal values[6] and the dihedral angle for the peptide bond was fixed in the trans conformation (180°). The sidechains for residues were represented as a point whose position was determined by averaging the sidechain centroids of the residues in 25 proteins in the Brookhaven Protein Data Bank (PDB). The only variables determining the conformation of each residue are thus the dihedral angles (phi and psi) defined by the bonds on either side of the alpha-carbon. For the current work, approximately 100 proteins in the PDB were used to generate probability plots of allowed phi-psi pairs (see example in Figure 5.1). Contours in the observed phi-psi distributions were then used directly to produce a set of randomly generated phi-psi pairs for each residue type. This set of phi-psi values was subsequently used to assign a conformation at random for each residue. With the exception of proline and glycine, the only significant

differences between the various residues was the relative populations of the alpha-helix and beta-sheet regions of their phi-psi probability plots (Table 5.1).

Empirical potentials describing the interaction between each amino acid pair were derived directly from the distribution of residues observed in known protein structures. Tanaka and Scheraga[2] and Crippen and Viswanadhan[7] have developed potentials with a similar basis in somewhat different ways. In the current work, the C_α-C_α distances of all residue pairs in 100 PDB proteins were tabulated and used to generate histograms of the number of occurences versus distance for each amino acid pair. The occurences in the 15-16Å range were used as a reference level since no specific contacts were assumed to occur at this large separation. The observed distributions of residue-residue distances were converted into free energies, assuming a Boltzmann distribution.

Potentials calculated in this way should include information about both long- and short-range packing interactions. The tendency of a residue to be buried can be inferred by examining the pair potentials for that residue with all other residues. Buried, hydrophobic residues consistently form many short-range contacts and show favorable interaction energies with a large number of other residues. Exposed residues on average form fewer close contacts (and thus have less favorable short-range interaction energies), although certain pairs may interact quite favorably. These tendencies can be demonstrated by comparing the pairwise interactions of leucine and lysine with other amino acids (Figure 5.2). Leucine, a commonly buried hydrophobic residue, interacts favorably with most amino acids at 4-10 Å. Lysine, on the other hand, shows much weaker interactions on average, despite strong tendencies to contact aspartate and glutamate residues. The potentials are affected by packing in addition to the general hydrophobic partitioning of amino acids. Since secondary structures generally associate in well-defined ways, they alter the potentials by favoring certain pairwise distances between residues. For example, two alpha-helices often align such that many of the distances from a residue in one helix to

a residue in the other helix fall in the range of 8-10 Å. Many of the contacts made by beta-strand residues are to residues on adjacent beta-strands and these are typically 4-6 Å long. Specific short-range interactions, such as electrostatic attraction or repulsion between charged sidechains, also perturb the potentials in a sequence-dependent way. The complete set of potentials is available as supplementary material or as part of the folding program. A thorough analysis of the potentials should reveal many of the patterns that have been found by previous surveys of the protein structure database. Our work, however, simply uses the potentials as an unbiased representation of the packing tendencies without attempting to explain them.

Since the potentials are based upon alpha carbon distances, they can include lower resolution structures from the PDB than those calculated by the method of Tanaka and Scheraga in which the position of all atoms must be well defined to determine if an interatomic contact has formed. To avoid the effect of near-neighbors along the sequence altering the distance potential without adding significant information about the conformation (since near-neighbors are covalently constrained to lie within a certain distance of each other), only the distances between residues separated by at least four other residues were included in the histograms. To calculate the energy of a given conformation, the distance between the i-th residue and all others, $j = 1$ to $i-5$ and $i+5$ to N, were calculated and the corresponding energies were totalled.

The interaction between the sidechains was initially ignored but early studies showed that this simplification was unreasonable. Using only alpha-carbon based potentials, the model peptide tended to pack into a globule much denser than the native protein. The space that should be occupied by sidechain atoms disappeared as the backbone potential was optimized. The pairing of two beta-sheets brings many backbone atoms in close contact with each other while maintaining good separation between the sidechains. Without explicit sidechains, the initial model brought many residues into this

potential well without allowing the subsequent addition of sidechains. As such, the model was modified to include potentials based on sidechain centroid-to-centroid distances. These potentials were calculated and applied as described above for the backbone potentials. The addition of these potentials significantly raised the average radius of gyration for the model protein, bringing it closer to that observed for the true structure.

A representative set of potentials is shown in Figure 5.2. For this and other potentials, the distribution appears to be the sum of two wells, one centered at 4-6Å (corresponding to residues in β-conformation) and another at 8-10Å (α-helical residues). Since conformation appeared to significantly affect the proper choice of a potential, the model was expanded to include separate potentials for each residue pair and residue conformation (based on phi-psi value) to determine the energy of a given structure. A single potential independent of residue type was applied to residue pairs separated by four residues (*i - i+4* and *i - i-4* interactions). This potential was determined using the procedure described above (a histogram of the distance between all *i - i+4* residue pairs, segregated by conformation, was generated and converted into a free energy function of distance). There are approximately 12,000 residues in the database used for this study and thus close to 12,000 *i - i+4* distances to be measured. If separate potentials for each amino acid pair were used, there would on average be less than 20 data points in the histogram for each potential. We therefore combined all *i - i+4* distances to produce a single potential which was used for all amino acid pairs.

Given a way to calculate the energy for any protein conformation, we required a means for adjusting the backbone angles to minimize the calculated energy. Simulated annealing is an optimization procedure by which high-energy or highly disordered systems may be 'cooled' in a controlled way that efficiently produces a system whose energy is close to that of the global minimum[8]. It has recently been used as a method for generating protein structures which satisfy specific interproton distance constraints[9]. Its ability to

rapidly sample over a highly convoluted energy surface makes it an attractive candidate as a protein conformation searching technique. As described by Kirkpatrick *et al.*, simulated annealing can be carried out as a Monte Carlo dynamics procedure in which the "temperature" of the system monotonically decreases during the course of the simulation[10]. In a run, the temperature is initially high such that the conformation of the system is rapidly changing and sampling a large volume of conformation space. As the run proceeds, the temperature is lowered slowly to allow the system to fall into a basin of attraction without becoming locked into the first local minimum reached. The specific heat of the system provides a good measure of the conformational transitions that are occurring and can be used to limit the rate of cooling to prevent premature freezing into local minima. The model described above was incorporated into a Monte Carlo dynamics program, written in FORTRAN, which was compiled and run on an FPS 264 [available in VAX and FPS versions, together with all the analysis programs used, upon request from CW]. Simulated annealing was carried out using a schedule for cooling similar to that suggested by Kirkpatrick *et al.* for the simulation of computer chip design[10]. During the run, a residue and a new phi-psi pair to replace the current value were chosen at random and the new conformation and corresponding energy were calculated. A thermal equilibrium was attempted at each temperature by applying the basic Metropolis algorithm[10]. In the Kirkpatrick annealing schedule, the relative temperature, kT, is lowered by a constant factor if either 1) 10 moves per site have been accepted since the last temperature change, or 2) 100 moves per site have been attempted without the required number of accepted moves. The first condition will be satisfied early during the run and allows the system to start at an arbitrarily high temperature and to rapidly cool down to a temperature at which true minimization can begin. Later, the second condition is satisfied and is used to signal the end of the run. The system is considered frozen and the run ended when the second condition is applied consecutively 3 - 10 times. For the current studies with crambin (46

residues), each annealled structure required approximately five to thirty minutes of CPU time on the FPS 264 (corresponding to approximately 1-6 minutes on the Cray X-MP and 1-6 hours on a VAX-780).

Knowledge of the secondary structure may be easily incorporated into the model by constraining the choice of dihedral angles to values within the region of phi-psi space corresponding to either alpha helix or to beta sheet. For simulations in which the secondary structure was considered known, the phi-psi values for alpha-helical residues were constrained to ($-140° < \Phi < -30°$, $-80° < \Psi < 30°$) and those for beta sheet residues were limited to ($-180° < \Phi < -30°$, $30° < \Psi < -160°$). [The probability of occurrence of phi-psi pairs was obtained from the observed phi-psi probability plots for each amino acid type.] Since this constraint limits the possible effects of altering the ordered residues (only a relatively small change in conformation is allowed), the program was written such that selection of these residues during the Monte Carlo procedure was slightly disfavored, typically chosen 60% as often as unassigned residues.

The structure of crambin has been solved to 0.945Å resolution with an R-value of 0.129, one of the most accurate protein structures ever determined[11,12]. The native protein contains a pair of alpha-helices, labelled here as A_1 (residues 7-19) and A_2 (residues 23-29), covalently joined to each other by a disulfide between residues 16 and 26. At either end of this helix pair are two short strands of beta sheet, B_1 (residues 1-4) and B_2 (residues 32-35), joined by a disulfide between residues 4 and 32 in an antiparallel sheet. The C-terminus (residues 36-46) has no regular structure although a reverse turn is defined by residues 41-44. [Secondary structure assignments were taken from the Protein Data Bank entry of the crambin coordinates (1CRN).] Cys-40 is covalently linked to Cys-3, thus bringing the terminus to lie adjacent to the first beta sheet. In simulations in which the secondary structure was considered known, the above assignments for helix and sheet were used.

Results and Discussion

Does simulated annealing solve the local minimum problem for a simplified protein model? Simulations using the simplified model described above were carried out to test the effectiveness of simulated annealing as a conformation searching technique. Fixed-temperature Monte Carlo simulations were done at several temperatures, starting from random conformations. The results were compared to a simulated annealing minimization in which the temperature was lowered periodically by 5% according to the schedule described above. Twenty structures were produced during each run and average characteristics for all simulations are shown in Table 5.2. The structures produced by the simulated annealing run had the lowest average energy and radius of gyration of structures from any of the simulations. The model peptide tended to fall into a compact structure while simultaneously avoiding the formation of bad overlaps. The contact maps for the structures can be averaged and those for some of the simulations are shown in Figure 5.3. For the non-annealed structures produced at low temperatures, contacts formed early between the residues at either terminus and the remainder of the peptide. These initial folds were sufficiently stabilizing that they could not be broken to allow more compact and lower-energy conformations to be made. At higher temperatures, these non-annealed structures failed to form many long-range contacts and most contacts occurred between residues separated by less than six residues. The annealing simulation contact map, however, showed that extensive contacts could be formed between pairs of helices and sheets to produce a low energy structure (Figure 5.4).

If the potentials used to calculate the energy of each conformation are an accurate approximation of the energetics for a true protein, the energy calculated for the native structure should be a global minimum, lower than all other possible conformations. For all

models attempted, however, several structures were found with calculated energies lower than that for the native. There are several possible explanations for this problem. A major simplifying assumption of our model is that the effect of solvent can be treated implicitly using potentials between pairs of amino acids, rather than by explicit interactions between individual residues and water molecules. Since most potentials are favorable in the 5-15Å range, the structures tend to be uniformly compressed to maximize the number of amino acid contacts that fall into this distance range. The average radius of gyration for all models was approximately 0.5Å smaller than that for the native structure. In several of the structures produced by annealing, the characteristic 'L'-shape of the native crambin (formed on one face by the helix pair and on the other by the beta-sheet) appear to be compressed into a 'V'-shape. This compression adds several helix-sheet interactions to the native helix-helix and sheet-sheet contacts, thus lowering the calculated energy relative to the native conformation. In reality, specific interactions of the solvent with the interface between the helix-pair and the beta sheet may be more favorable than those formed when the protein is compressed. Unless solvent is included explicitly, however, there is no obvious way to prevent this compression without removing the driving force for folding. Other possible inaccuracies in the potentials that have been used will be discussed later.

There are two basic requirements for any simulated folding process: 1) random starting conformations should converge to a single minimized structure, and 2) the structure of the model at its global minimum should be identical to the native, experimentally observed structure. In practice, these goals are related and they have never been completely satisfied by a computer model. If random starting structures do not converge, it is unlikely that the global minimum is one of the structures in the final set. Without the global minimum defined by the model to test against the native structure, it is impossible to determine the accuracy of the model. To assess the ability of the annealing procedure to converge to a single minimum (self-convergence) and the similarity of the final set of

structures to the native protein, the minimized structures were compared using two similarity metrics. The most common measure for comparing protein crystal structures, the r.m.s. deviation obtained by optimally translating and rotating one structure onto another (RMS), has questionable usefulness when comparing two structures that are similar to low resolution[13]. Another common measure, the r.m.s. distance matrix error (DME), will determine the extent to which a pair of structures have similar patterns of contacts but ignores any differences in the chirality of the structures[14]. For instance, a left-handed and a right-handed helix may have a DME deviation approaching zero but a relatively high RMS deviation. The DME deviation is typically less than the RMS deviation by 30%. Both metrics were determined for the structures produced by the different minimization procedures (Table 5.2). Figure 5.5 shows several sample model structures produced by simulated annealing, together with the crystal structure. One of these structures (fifth row) is especially interesting: while its general fold appears correct at first glance, the RMS deviation is quite large (7.5 Å). The DME for this structure, on the other hand, is only 3.5 Å. Crambin is 'L-shaped,' with the paired alpha-helices and the beta sheet lying perpendicular to each other. The model structure has the same L-shape but is inverted. The DME metric, unlike the RMS metric, is able to identify the general correctness of this fold (correct in the sense that the long-range contacts which energetically favor this conformation have formed).

To test the first criterion, self-convergence was estimated by calculating the average deviation between all pairs of structures from a given run. The deviation between structures produced by simulated annealing averaged 4.21 Å (DME), signficantly less than that for all of the fixed temperature Monte Carlo simulations (4.94 Å - 7.91 Å). To estimate the ability to converge to the true global minimum, the deviation of the model structures from the crystal structure was calculated. The average deviation of the annealed structures from the native protein is 7.58 Å (RMS), 4.76 Å (DME). [The model closest to

the native structure had a deviation of 4.01 Å (RMS), 3.17 Å (DME).] The deviation between any single structure and the crystal structure is thus only slightly more than the deviation between any pair of model structures. It is possible, therefore, that the inability to produce structures closer to the native is limited by the minimization procedure (a convergence problem), rather than by the potentials and the model. However, since the native state is not a global minimum as defined by the potentials, it will be impossible to obtain the correct structure by a complete search of conformation space. A contact map for the collection of minimized structures can be generated by averaging the individual contact maps for each structure. The distance matrix error of the average simulated structure with the native is 3.76 Å, significantly lower than the average error of any single structure with the native (4.76 Å). This suggests that while every structure differs in some way from the native, the average tends to fold in a way quite similar to the native. We are currently using distance geometry methods to rebuild a three-dimensional structure from the average contact map.

Does minimization produce 'protein-like' conformations? The models produced by simulated annealing were analyzed in a variety of ways to determine if they are structurally similar to true proteins and to the crambin structure in particular. Hagler and Honig, in their critique of early protein-folding models, emphasize the need to examine the structural details of the model protein to determine the accuracy of the simulation. For the current work aspects of secondary structure, the hydrophobic effect, and disulfide bond formation were considered. As an additional test of the folding procedure, the crambin sequence was randomly shuffled to yield a sequence of identical composition but with no homology to crambin. Differences between minimized structures produced for the native and shuffled sequences can then be used to indicate the sequence-dependence of the modelling.

Secondary structure

Inspection of the phi-psi plots for amino acids shows that each have different biases towards alpha-helical or beta-sheet conformation. These biases can be used in a simplistic way to predict secondary structure for a sequence. Assuming that the choice of phi-psi value for a residue is independent of all other neighboring residues, the probability of having a regular structure spanning a window of residues can be calculated as the product of the probabilities of forming that structure for each residue in the window. [A regular structure is defined here as a linear set of residues whose phi-psi values are all within the same region of phi-psi space.] The probability, P_i^a, that the *i*-th residue will lie in a type-*a* regular structure of length equal to or greater than *w*, may be calculated as:

$$P_i^a = \sum_{n=i-w+1}^{i} \prod_{m=n}^{n+w-1} p_m^a$$

where p_m^a is the probability that the *m*-th residue will have a phi-psi value corresponding to type-*a* regular structure. [This procedure is similar to the Chou-Fasman secondary structure algorithm[15], although the p_m^a's which define the probabilities are determined quite differently.] This method was applied to the crambin sequence over a window of four residues and identified several regions with slight alpha-helix or beta-sheet tendencies (Figure 5.6). The tendencies calculated by this method are equivalent to those observed for the starting structures (the profiles obtained from non-minimized structures were identical to those in Figure 5.6). When simulated annealing was carried out with no assigned secondary structure, the observed regular structures occurred in the same regions as in the native structure and much more frequently than that predicted by the simple method described above. The tendency to form regular structures during the minimization seems to be initiated by a bias in the phi-psi values but enhanced during the simulated folding.

Reasons for the enhancement of secondary structure are unclear. One possibility is that the formation of secondary structure allows residues to pack in defined ways and thus tends to increase the number of favorable contacts and lower the total energy. However, there was a wide range in the amount of regular structure formed in the minimized models (from 0 to 45% of all residues) and no correlation to the number of contacts or to the calculated energy could be found. Several models with no regular structure were produced with an equal number of contacts and energy as those models with large amounts of regular structure. Thus, while regular structures may be enhanced during folding, they cannot be said to be required for folding. A reasonable explanation for the enhancement of secondary structure is that the potentials favor certain pairwise distances (4-6Å, 8-10Å) and that one way (but not the only way) to produce folded structures with these distances is to form secondary structure (β-strand pairing creates residue-residue distances of 4-6Å, α-helix pairing forms 8-10Å contacts).

Prior to minimization, there are several regions that show tendencies towards both helix and sheet. After minimization, there is little overlap between the two types of structure, making it much easier to assign secondary structure to the sequence. A variety of prediction schemes have been developed that use local sequence information to assign secondary structure for each residue[23,16]. None of these techniques, however, is more than 70-80% accurate when tested against a collection of known structures. The inability of these methods to correctly predict secondary structure has been ascribed to long-range interactions which influence secondary structure formation but cannot be determined from a simple survey of the sequence[17]. By deleting segments of the sequence and attempting to fold a partial protein, one can estimate the importance of these long-range effects. Decapeptides corresponding to regions of the crambin sequence were folded using the model aand the resulting pattern of secondary structure was compared to that obtained when the entire protein was folded (Figure 5.7). There are large differences between the

secondary structure of a decapeptide in solution and that of a decapeptide in the context of the protein. These results demonstrate convincingly that the formation of secondary structure during the simulation requires information from a sequence as large as the protein itself. Attempting to model tertiary structure may take long-range interactions into account and thus improve the prediction of secondary structure.

The formation of the correct secondary structure would not be surprising if the starting conformations had this native structure built into them. The simplified models of both Levitt-Warshell and Hagler-Honig assume a starting conformation in which the BPTI peptide is completely extended. Since the folded BPTI structure consists mostly of extended β-strands, separated by turns, it is not surprising that the simulations do as well as they do. In contrast, our model makes no assumptions about the starting structure as phi-psi values for each residue are chosen at random from a phi-psi probability plot. Any starting structure forced onto the model is quickly lost because of the randomizing effect of high temperature in the initial stages of refinement. This can be shown by comparing simulations in which either a completely alpha-helical structure or a completely extended (β–strand-like) structure is used as the starting point for refinement. No significant changes in the observed patterns of secondary structure or the formation of specific contacts are observed for both simulations.

Do residues pack correctly?

A major problem in the calculation of protein structure involves the interaction with solvent. Novotny *et al.* showed that *in vacuo* molecular mechanics applied to a completely incorrect starting conformation was able to produce a minimally altered structure with a potential energy comparable to the true structure[18]. While their final structure had no bad van der Waals' contacts or strained covalent interactions, the distribution of polar and non-polar residues at the protein surface and interior was obviously skewed and thus affected

the electrostatic contribution to the enthalpy when solvent was taken into account. A correct globular structure should be expected to bury a significant fraction of its hydrophobic residues and to expose most charged residues. To determine if the model behaved in this way, an algorithm developed by Ponder and Richards[19] was used to define which residues are buried for each structure produced by the simulation. The Kite-Doolitle index[20] was used to calculate the hydropathy of buried and exposed residues. The average hydropathy of buried residues was 0.69, significantly higher (indicating greater hydrophobicity) than that for the non-buried residues (0.10). The tendency to optimize hydrophobic forces by burying nonpolar residues and excluding highly polar residues appears to be well satisfied for these structures. The corresponding values for the native protein, 0.40 for buried residues and 0.33 for exposed residues, indicated less of a tendency for a hydrophobic core to form in the native protein than in the simulated models. Crambin has unusual water solubility for a protein and forms crystals which are surprisingly non-hydrated[18]. The average hydropathy of buried and exposed residues in several other proteins was calculated to determine if the typical model structure or the native crambin structure is unusual with respect to partitioning of hydrophobic and hydrophillic residues. The average difference between buried and exposed residues in basic pancreatic trypsin inhibitor (5PTI), cobra venom toxin (1CTX), staphylococcal nuclease (2SNS), and alpha-lytic protease (2ALP) was calculated using the PDB structures for these proteins (data not shown). The mean difference in hydropathy between buried and exposed residues for these test proteins was 1.30, significantly higher than that for the both native and model crambins. It appears that the simulation is attempting to force the formation of a hydrophobic core for a protein which does not require it. The potentials used for the simulation may not be appropriate for a protein such as crambin whose structural properties differ significantly from those of the basis set which originally defined the potentials. The fact that several structures could be found with a calculated energy lower than that of the

native may be due to the incorrect choice of potentials for a protein that does not enormously favor the formation of a hydrophobic core. The stability of crambin may be due largely to disulfide bond formation rather than optimization of the hydrophobic effect. The ability of the model to simulate disulfide bond formation will be discussed later.

Formation of specific contacts

The time during the simulation when the final contacts first occur was tabulated for each structure and used to calculate the average age of each contact. The largest contact region, that between the beta-sheet strands, was usually formed towards the end of the simulation. This is not surprising since the two helices to which the strands are attached must move together before allowing the strands to be closely paired in an antiparallel sheet. The oldest contacts, as expected, form between residues close to and on opposite sides of each of the turn regions. These contacts are those mostly likely to be associated with a change in conformation which will bring large numbers of residues in contact with one another. The average contact map shows many features of the native, *e.g.* specific contacts between the two beta-sheets, helix-helix pairing at well-defined sites, disulfide formation between the N- and C-termini, etc.. The map is inaccurate in that the position of the unordered C-terminus is averaged somewhat over several positions. This may be expected since the lack of assigned secondary structure for the 11 C-terminal residues allows them considerably more conformational flexibility than for the sequences of assigned secondary structure.

Effect of disulfide formation on the folding pathway

The potential for the interaction between two cysteine residues is about twice as deep as the average potential. One way that inaccuracies in the average contact map could occur would be if the formation of a disulfide locked the conformation of intervening

residues and prevented their proper minimization. The formation of native and non-native disulfides was compared for the 100 structures produced by simulated annealing. The native protein is cross-linked by disulfides between residues 16-26, 4-32, and 3-40. Defined in terms of alpha-carbon separation (less than 6Å), a native disulfide also exists between residues 3-32. During the folding simulation, the native disulfides occurred more often than any of the non-native disulfides although they occurred only 0.32 times on average. Non-native disulfides occurred much less frequently, an average 0.035 times. The average age of all disulfide contacts (15.8% of the total simulation length) is slightly more than the average for all non-disulfide contacts (13.5%). If disulfide formation acts as a steep local minimum to prevent the continued searching of other conformations, one would expect the disulfide contacts to be among the last contacts formed, significantly younger than the non-disulfides.

One way to test the role of disulfides in the simulated folding process is to alter the cys-cys potential. By substituting the potential for cys-cys residues with one calculated for all residues (generalized potential), the number of both native (0.138/structure) and non-native disulfides (0.015/structure) was signficantly decreased. 80% of these native disulfides were due to links between residues 3-32 and 4-32, likely due to the natural tendency for beta sheets to pair. The cys-cys potential was also replaced by one based on the distribution of disulfide distances rather than cys-cys distances. Since no disulfide bonds are formed at 15-16Å alpha carbon separation, this potential was centered at an arbitrary low level (-20kT) with zero energy for distances outside of the distribution. Replacing native disulfides (3-40,4-32,16-26) with a disulfide-specific potential effectively selected for the formation of native disulfides over non-native (0.50 vs. 0.05), although their number did not increase dramatically. Explicit disulfides did adjust the relative occurence of certain native contacts, increasing the fraction of structures with contacts in the A_1-A_2 and B_1-R regions, but did little to change the overall pattern of contact

formation. Deepening the disulfide potential for native disulfide pairs apparently does not lead to premature freezing of the conformation as judged by the lack of non-native contacts.

Structures produced by shuffled and degenerate sequences

The structure of a protein is obviously sensitive to its sequence. A model which generates the same structure regardless of sequence is thus certainly incorrect. A simple test of the current model is to randomly shuffle the crambin sequence, maintaining composition, and to repeat the simulation. The shuffled sequence (Figure 5.8) was initially annealed with no secondary structure assigned. Secondary structure formed in a well-defined way, but the pattern of helices and sheets was completely different from that of the native sequence. With this secondary structure assigned, the final average contact map is entirely different from that produced by the native sequence. Even with the correct native secondary structure assignment, the contact map produced by the shuffled sequence did not resemble the native sequence-based model. These observations support the hypothesis that the model using the native sequence is mimicking crambin's structure in a sequence-dependent way.

Results for other proteins

Preliminary work has been done to simulate the folding of proteins other than crambin. Bovine pancreatic trypsin inhibitor has been used as a standard for many protein folding schemes because of its small size and experimental information on its folding pathways[21]. One of the structures produced by simulated annealing with the current model has a DME deviation of 4.5 Å from the crystal structure. Levitt and Warshel[1] and Kuntz *et al* [3] have used other approaches to simulate the folding of BPTI. The best structures generated by their methods have a DME deviation of 5.3 Å and 4.7 Å respectively, slightly worse than that generated by the current model using simulated annealing. While this

improvement is small, it is important to note the fundamental differences between the starting assumptions of other models and our own. As mentioned before, the Levitt-Warshel model assumes certain properties of the starting conformation and the energetics that would force the formation of a BPTI-like structure. Kuntz et al. use a model with several adjustable parameters which are fitted to optimize the agreement to the BPTI crystal structure. In addition, they assume that the three disulfides found in the crystal structure are known constraints before the minimization. In contrast, our model is parameterized completely by the database of known crystal structures and assumes nothing about the BPTI structure beyond the sequence of amino acids. It is therefore quite surprising that the current model appears slightly better than other models which include BPTI-specifc constraints. The improvement suggests that the specific constraints imposed in these earlier models can be replaced in a generalizable way by potentials derived from a large database of observations. Subsequent work shall determine whether the annealing procedure yields intermediate structures which are similar to those inferred experimentally from data on disulfide formation. If such intermediate structures exist, one may conclude that the annealing process is reproducing the dynamics of the folding process, independent of its ability to converge to the native minimum structure.

Conclusions

A simplified representation of proteins is presented in which the conformation of the backbone and sidechains is specified by the dihedral angels for each residue. An empirical energy is calculated using distance-dependent potentials between each pair of amino acids. These potentials have been derived from the distribution of pairwise distances observed in known crystal structures. Simulated annealing is used as a refinement

technique to sample many different conformations and minimize the energy of the model structure. Several results have been obtained using the crambin sequence to test the model:

1. Starting from random conformations, secondary structure forms in the model peptide where it occurs in the true structure.
2. The formation of secondary structure is sequence-specific and depends upon the context of the sequence (i.e. is influenced by long-range interactions).
3. With the secondary structure assigned, alpha-helices and beta-strands associate as they do in the true structure.
4. Certain pairs of cysteines, corresponding to native disulfides, associate much more frequently than other cysteine pairs. This pairing is partially driven by the folding of the rest of the protein, and partially driven by the strong attractive potential between two cysteines.
5. The empirical potentials favor the formation of a hydrophobic core of residues. This partitioning is imposed during the crambin simulation, even though the true crambin structure lacks a hydrophobic core.

One problem with the empirical approach used by this model is that the true energetics which drive protein folding are deeply embedded in the amino acid pair potentials. One might ask, for instance, why does a helix form in crambin between residues 22 and 29? The reason cannot be attributed by this model to particular hydrogen bonds, backbone conformational preferences, etc., since the simulation depends on the entire collection of empirical potentials assigned to these and other amino acids. Since, however, minimization using the empirical potentials successfully predicts a helix in this region, the problem has been reduced to explaining how specific interactions within the database of structures give rise to the potentials in the first place.

Table 5.1: Phi-psi probabilities

Amino Acid	Distribution					
	Alpha		Beta		Neither	
	#	%	#	%	#	%
Alanine	567	54.5	312	45.0	162	15.6
Arginine	174	48.5	130	36.2	55	15.3
Asparagine	224	41.1	168	30.8	153	28.1
Aspartate	324	45.2	214	29.8	179	25.0
Cysteine	111	39.1	121	42.6	52	18.3
Glutamine	198	45.8	162	37.5	72	16.7
Glutamate	353	58.8	129	21.5	118	19.7
Glycine	242	21.5	183	16.3	701	62.3
Histidine	103	40.2	85	33.2	68	26.6
Isoleucine	234	39.1	268	44.7	97	16.2
Leucine	420	47.4	318	35.9	148	16.7
Lysine	420	51.7	205	25.2	188	23.1
Methionine	88	47.8	63	34.2	33	17.9
Phenylalanine	187	41.2	183	40.3	84	18.5
Proline	178	35.6	244	48.8	78	15.6
Serine	414	42.9	377	39.0	175	18.1
Threonine	299	39.7	333	44.2	121	16.1
Tryptophan	78	43.1	73	40.3	30	16.6
Tyrosine	148	35.1	211	50.0	63	14.9
Valine	316	35.7	415	46.8	155	17.5

Phi-psi probability plots were calculated for each amino acid type as described in METHODS. Phi-psi space was broken into three regions termed alpha ($-140° < \Phi < -30°, -80° < \Psi < 30°$), beta ($-180° < \Phi < -30°, \Psi < -160°$ or $\Psi > 30°$), or neither (all remaining areas). These labels are for reference only and do not indicate that all phi-psi pairs within each region were obtained for a particular secondary structure (i.e. alpha helix or beta sheet). The table lists the number of occurences and fraction of residues found in each region of phi-psi space for each amino acid type.

Table 5.2: Characteristics of fixed temperature and annealed simulations

Temperature	Energy (kT)	Rad. of gyr (Å)	Total contacts	RMS Native (Å)	DME Native (Å)	DME Internal (Å)	DME Average (Å)
0.5	2531.6	9.68	781.2	8.15	5.34	5.03	4.06
1.0	2507.2	9.90	755.6	7.95	5.54	5.03	4.30
2.0	2528.5	9.65	784.6	8.01	5.36	4.94	4.14
5.0	2529.8	9.77	754.9	7.72	5.16	4.95	3.86
10.0	2509.7	9.99	727.0	8.36	5.51	5.35	4.08
20.0	2690.8	12.24	590.5	8.34	7.01	7.04	5.29
100.0	3357.3	17.47	469.7	12.42	14.28	7.91	13.46
sim. anneal	2349.7	9.15	828.1	7.54	4.76	4.21	3.76
crystal structure	3192.9	9.70	876.0	----	----	----	----

Monte Carlo dynamics was run at the indicated fixed temperatures. For the simulated annealing run, the temperature was initially set to 100, and gradually lowered. RMS native refers to the average RMS deviation of each model structure from the PDB crystal structure. DME indicates the average distance matrix errors between each model structure and the crystal structure (native) and between each model structure and every other structure (internal). The average DME is the distance matrix error calculated by comparing the contact map of the crystal structure to the average contact map generated from all the model structures of a given simulation.

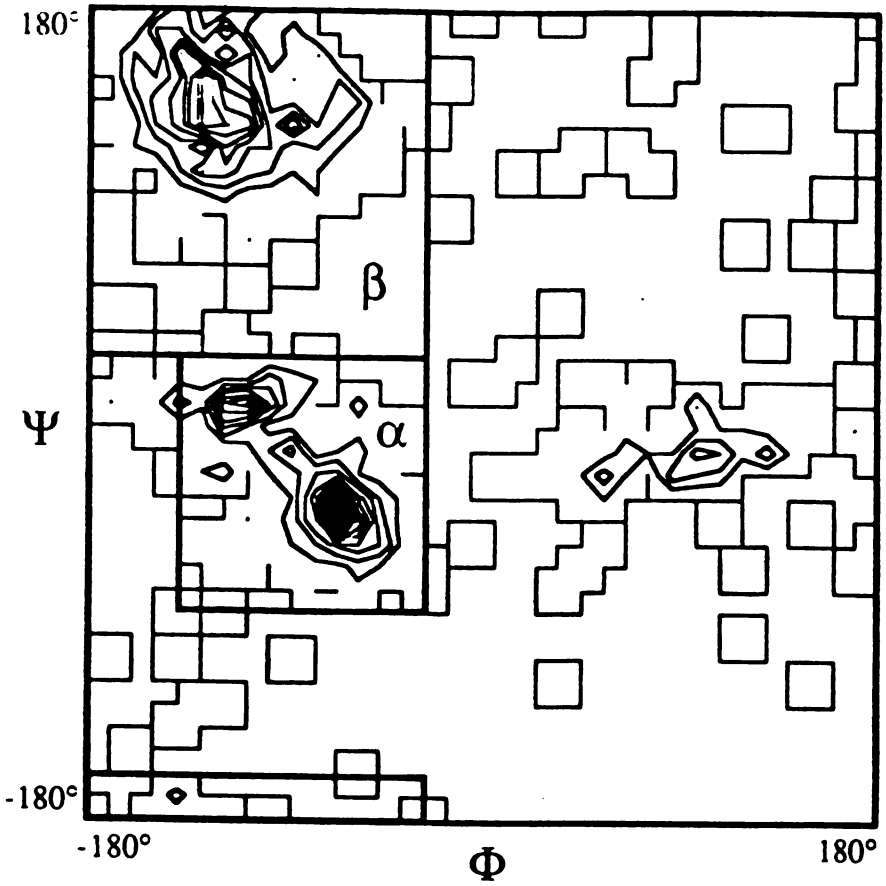


Figure 5.1: Phi-psi probability plot for valine. Phi-psi probability plots were calculated for each amino acid as described in METHODS. The probability of a residue having a set of phi-psi values was calculated for each 10° x 10° pixel of phi-psi space. Probability contours are drawn for the valine plot in intervals of 0.0015 (twice the average value for all pixels). The regions defined as alpha and beta conformations are enclosed and labelled.

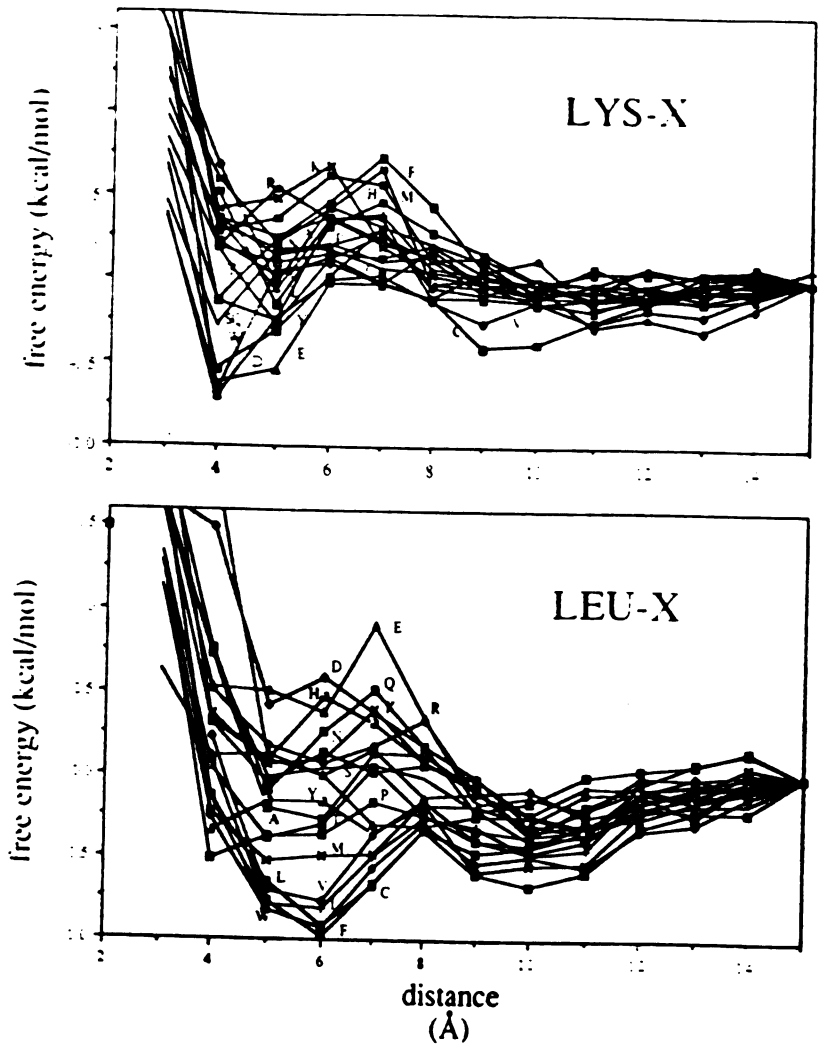


Figure 5.2: Potential for interaction with leucine and lysine

Upper panel: The sidechain distances from every lysine to every other amino acid sidechain were calculated for 100 PDB structures and used to produce a distance-dependent potential for the Lys-X interaction as described in METHODS. Each potential is labelled by the corresponding one letter code for the amino acid. Negatively charged and other polar amino acids (D,E,W,S,Y) are strongly attracted to lysine, while positively charged amino acids (R,H,K) and some hydrophobic amino acids (F,M) are repelled.

Lower panel: As above but for the Leu-X interactions. As expected, the hydrophobic amino acids (F,W,C,I,L,V,M) interact favorably with leucine while the charged amino acids (E,D,H,K,R) interact with it poorly.

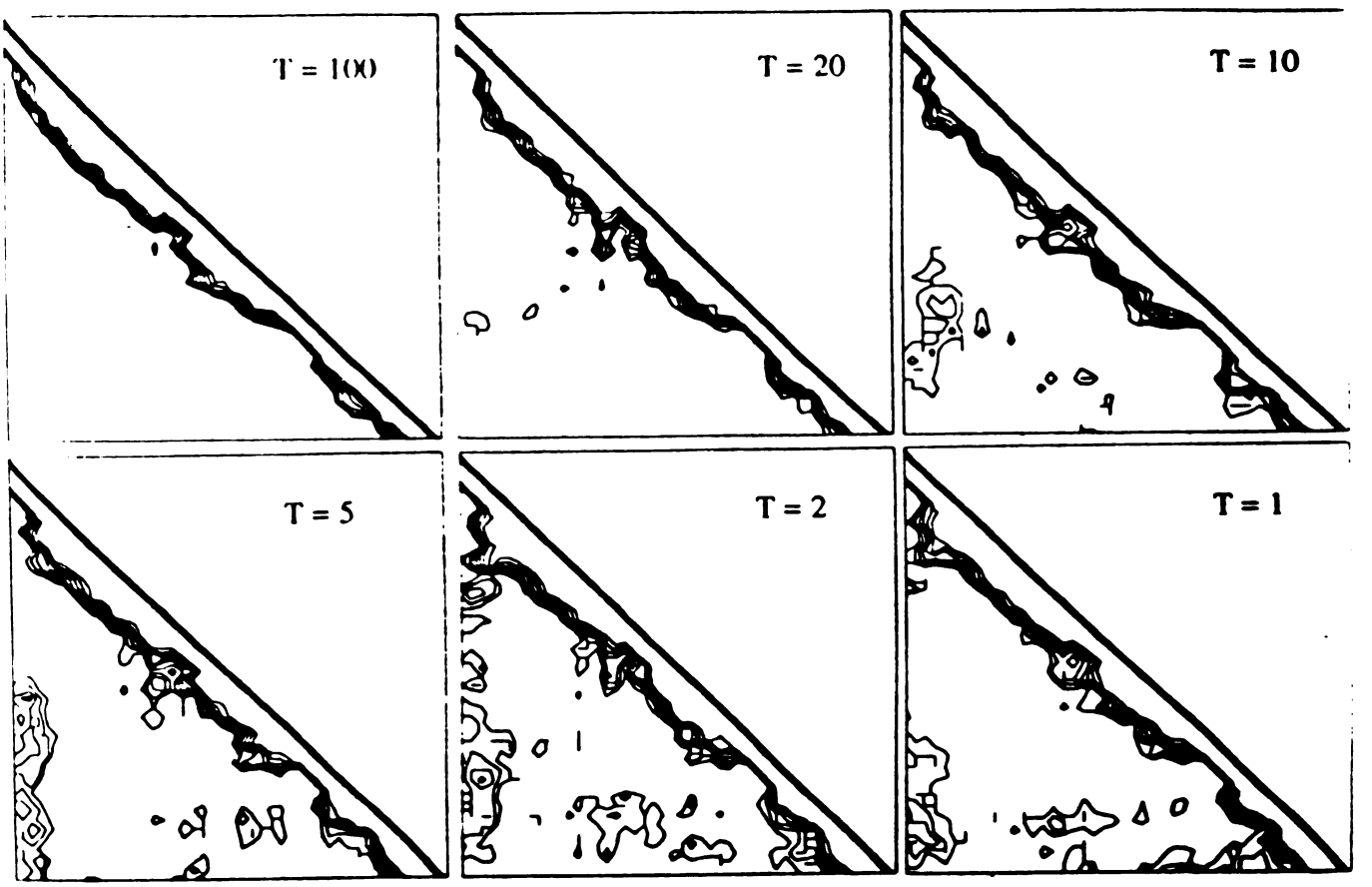


Figure 5.3: Average contact maps for simulations as a function of temperature. Average contact maps were generated for structures from several of the fixed-temperature (non-annealed) simulations. In a contact map the protein sequence runs from the N- to C-terminus down the vertical axis and left-to-right along the horizontal axis. Contacts were defined for each pair of residues whose alpha-carbon atoms were seperated by less than 10 Å. The contour levels in the plots indicate the frequency with which contacts between residues at the intersection of each point occurred among the structures produced by each simulation.

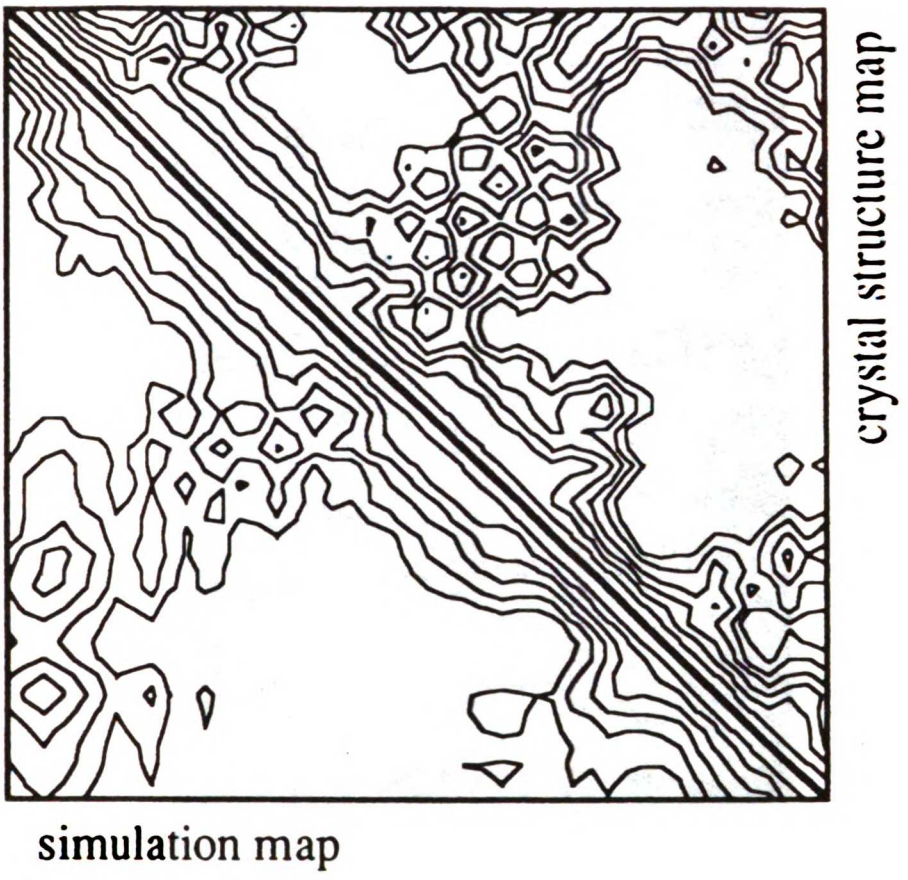


Figure 5.4: Average contact map for the annealed model and the native structure

The average contact map calculated for the 100 structures produced by simulated annealing with explicit disulfides is shown in the lower triangle while the contact map for the native structure is shown in the upper triangle. The contours for both contact maps are based on the distance between alpha carbons for each residue (with the shortest contacts enclosed by the highest contours).

Figure 5.5: Several minimized model structures and the native structure

A variety of model structures produced by minimization are shown in the same orientation as the crystal structure. (a-helix = red, b-strand = blue, reverse turn = green).

First row: The crystal structure (PDB name 1CRN)

Second row: The "best" structure produced by minimization (RMS = 4.01Å, DME = 3.17Å). The alignment of helices is nearly perfect (r.m.s. deviation < 1.5Å) and the overall construction of the model is correct, although the beta strands are poorly packed.

Third row: The starting structure which yielded the best structure on minimization.

Fourth row: The "worst" structure produced by minimization (RMS = 9.59Å, DME = 5.13Å). The C-terminal 15 residues have packed against the open face of the helix pair rather than lying next to the beta sheet as they do in the native structure.

Fifth row: An unusual structure produced by minimization. The RMS deviation of the minimized structure from the native structure (7.5Å) would not suggest that the predicted fold is similar to the true fold. The low DME deviation (3.5Å), however, suggests that the model structure may have a pattern of contacts highly similar to the native. Inspecting the structures, it is clear that the overall shape and local contacts in the native structure have been reproduced by the model but that the two helices have been swapped.

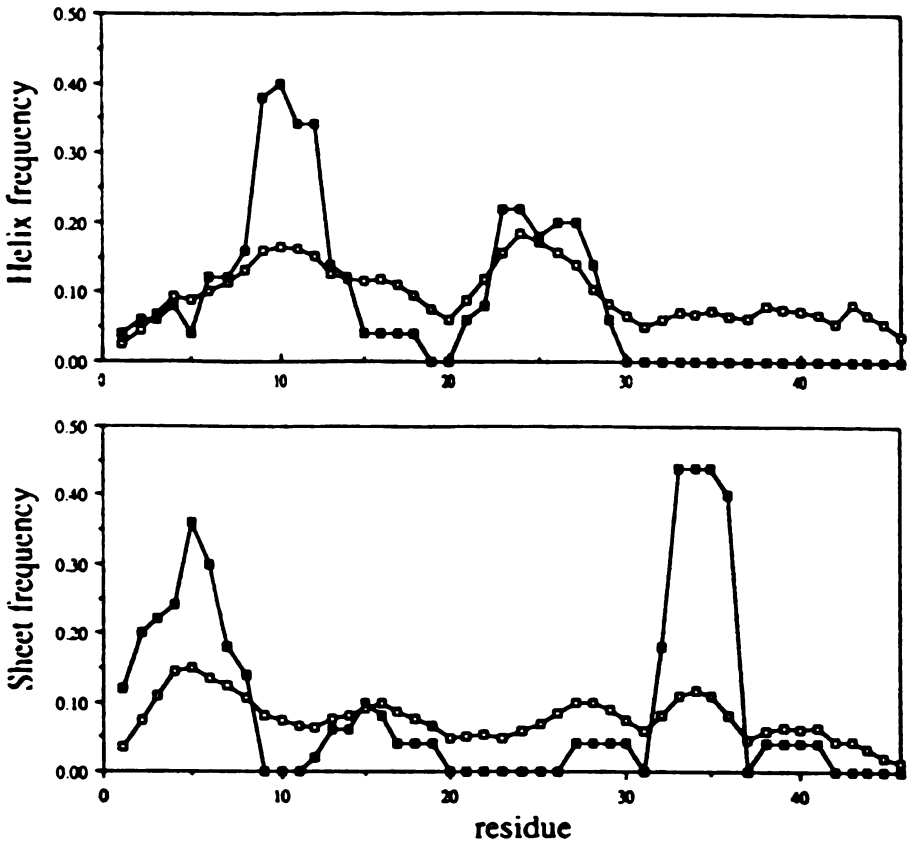


Figure 5.6: Formation of secondary structure

The occurrence of regular structure, defined as four or more residues in either alpha-helical or beta-sheet conformation, was tabulated for non-minimized starting structures (open squares), and for 50 structures produced by simulated annealing with no assigned secondary structure (solid squares). The formation of helix is limited to those regions in which significant bias for helix formation exists, corresponding, in fact, to the helical regions in the native structure (residues 7-19, 23-29). The formation of beta sheet is significantly enhanced during minimization in regions overlapping the sequences of beta sheet in the native structure (residues 1-4, 32-35).

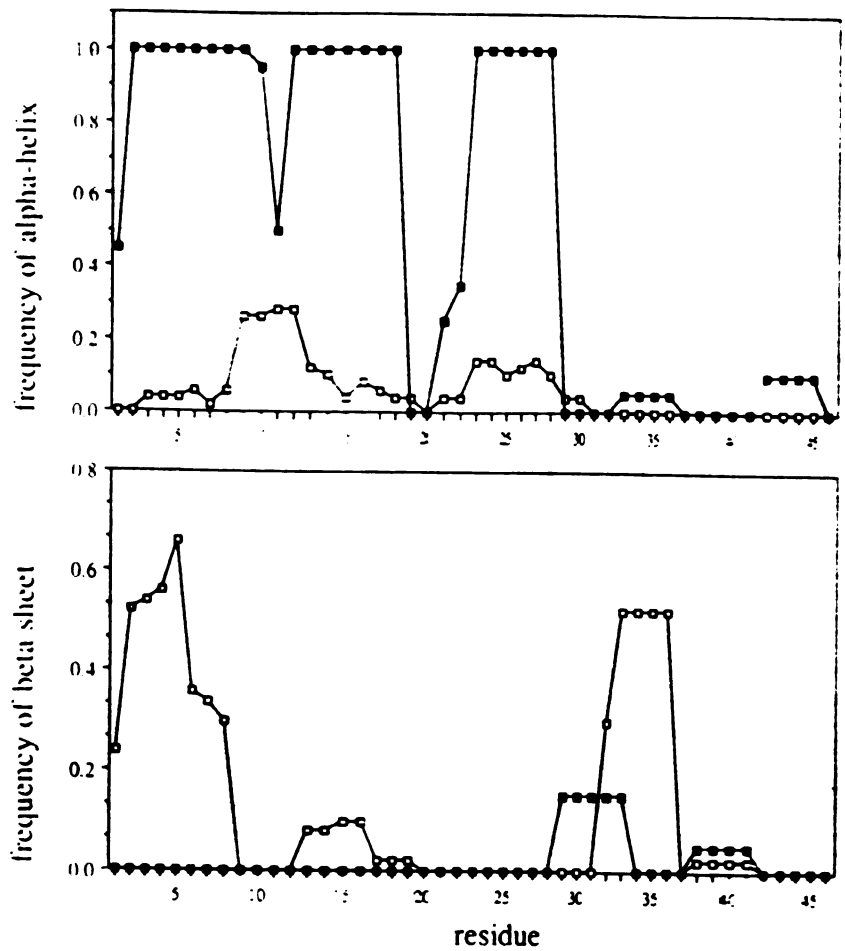


Figure 5.7: Formation of secondary structure by peptide fragments

The effect of sequence context on the formation of secondary structure was determined by breaking the crambin sequence into a series of non-overlapping decapeptides. Twenty structures were generated by minimizing each decapeptide as described in the text. The resulting secondary structure (solid squares) is compared to that calculated for the intact crambin minimization (open squares). The frequency of helix formation is significantly increased while beta strands are virtually eliminated.

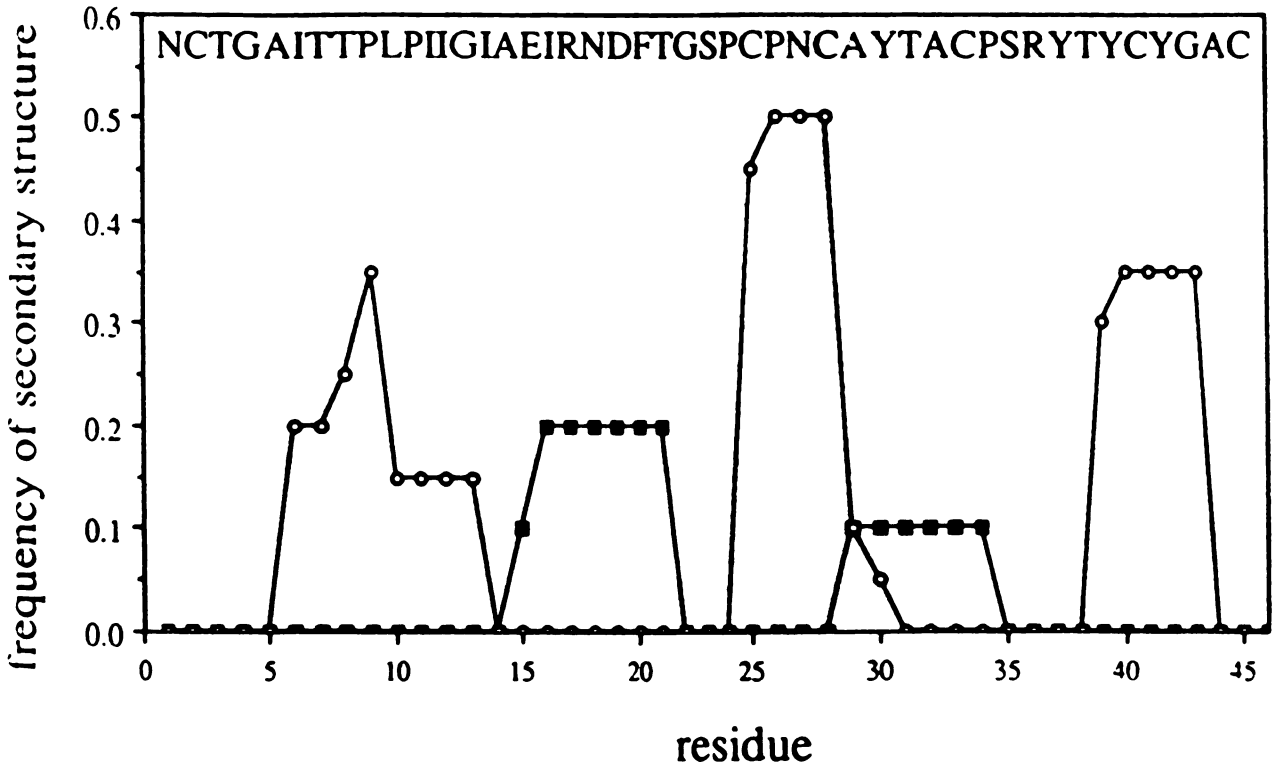


Figure 5.8: Results with a shuffled sequence

The above randomly-generated sequence, with the same composition as crambin but no homology to it, was used to test the sequence dependence of the modelling. The formation of secondary structure was calculated as described in figure 5.6 (open squares indicate alpha-helical structure while solid squares indicate beta-sheet structure).

References

[1]Levitt, M., Warshel, A. "Computer simulation of protein folding." Nature. 253:694-698, 1975.

[2] Tanaka, S., Scheraga, H.A. "Medium- and long-range interactions parameters between amino acids for predicting three-dimensional structures of proteins." Macromolecules. 9(6):945-950, 1976.

[3] Kuntz, I.D., Crippen, G.M., Kollman, P.A., Kimelman, D. "Calculation of protein tertiary structure." J. Mol. Biol. 106:983-994, 1976.

[4] Hagler, A.T., and Honig, B. "On the formation of protein tertiary structure on a computer." Proc. Natl. Acad. Sci. USA. 75(2):554-558, 1978.

[5] Go, N., Taketomi, H. "Respective roles of short- and long-range interactions in protein folding." Proc. Natl. Acad. Sci. USA. 75(2):559-563, 1978.

[6] Ramachandran, G.N., Sasisekharan, V. "Conformation of polypeptides and proteins." Adv. Protein Chemistry. 23:283-437, 1968.

[7] Crippen, G.M., Viswanadhan, V.N. "Sidechain and backbone potential function for conformational analysis of proteins." Int. J. Peptide Protein Res. 25:487-509, 1985.

[8] Kirkpatrick, S., Gelatt, C.D., Jr., Vecchi, M.P. "Optimization by simulated annealing." Science. 220:671-680, 1983.

[9] Nigles, M., Gronenborn, A.M., Brunger, A.T., Clore, G.M. "Determination of three-dimensional structures of proteins by simulated annealing with interproton distance constraints. Application to crambin, potato carboxypeptidase inhibitor, and barley serine proteinase inhibitor 2." Protein Engineering. 2:27-38, 1988.

[10] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. " Equation of state calculations by fast computing machines." J. Chem. Phys. 21:1087-1092, 1953.

11 Hendrickson, W.A., Teeter, M.M. "Structure of the hydrophobic protein crambin determined directly from the anomolous scattering of sulphur." Nature. 290:107-113, 1981.

12 Teeter, M.M. "Water structure of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in crystals of crambin." Proc. Natl. Acad. Sci. USA. 81:6014-6018, 1984.

13 Cohen, F.E., Sternberg, M.J.E. "On the prediction of protein structure: The significance of the root mean square deviation." J. Mol. Biol. 138:321-333, 1980.

14 Havel, T.F., Kuntz, I.D., Crippen, G.M. "The theory and practice of distance geometry." Bulletin of Mathematical Biology. 45:665-720, 1983.

15 Chou, P.Y., Fasman, G.D. "Prediction of protein conformation." Biochemistry. 13(2):222-245, 1974.

16 Cohen, F.E., Abarbanal, R.M., Kuntz, I.D., Fletterick, R.J. "Secondary structure assignment for alpha/beta proteins by a combinatorial approach." Biochemistry. 22:4894-4904, 1983.

17 Schulz, G.E., Schirmer, R.H. "Principles of Protein Structure." Springer-Verlag, New York, 1979.

18 Novotny, J., Bruccoleri, R., Karplus, M. "An analysis of incorrectly folded protein models: implications for structure predictions." J. Mol. Biol. 177:787-818, 1984.

19 Ponder, J.W., Richards, F.M. "Tertiary template for proteins-use of packing criteria in the enumeration of allowed sequences for different structural classes." J. Mol. Biol. 193:775-791, 1987.

20 Kite,J., Doolittle,R.F. "A simple method for displaying the hydropathic character of a protein." J. Mol. Biol. 157:105-132, 1982.

21 Creighton, T.E. "Experimental studies of protein folding and unfolding." Prog. Biophys. Mol. Biol. 33:231-297, 1978.

Chapter 6 :

A predicted structure for the apolipoprotein-binding domain of the LDL receptor

Introduction

By directing the specific endocytosis of cholesterol-rich apolipoproteins, the low density lipoprotein (LDL) receptor plays a key role in mediating the cellular uptake of lipid and cholesterol. The receptor functions by binding with very high affinity ($K_d \approx 10^{-9}$-10^{-10} M) to both apolipoprotein-E and apolipoprotein-B, proteins found on the surface of very low density lipoprotein (VLDL), chylomicron remnants, high density lipoprotein (HDL), and low density lipoprotein (LDL) (Pitas, 1980). Following binding, the LDL receptor together with its associated lipoprotein particle is internalized into clathrin-coated vesicles, allowing the intracellular utilization of the lipoprotein's lipid and cholesterol (reviewed by Brown, 1980). Interaction of the LDL receptor with its protein ligands thus acts to tightly regulate the circulating concentrations of various lipoproteins, including LDL. Given the positive correlation between LDL levels and risk for coronary heart disease, this receptor-ligand interaction is of major clinical relevance. We have thus initiated crystallographic studies of the LDL receptor-binding domain of human apolipoprotein-E, and have recently determined the three-dimensional structure of this protein (Wilson, 1991a). We now describe our efforts to predict the structure of the apolipoprotein-binding domain of the LDL receptor and to model the interaction between the receptor and one if its major ligands.

The primary sequence of the LDL receptor indicates several different domains, including seven copies of a short cysteine-rich motif, three copies of an epidermal growth factor-like domain, a glycosylated linker region, a short transmembrane peptide, and a cytoplasmic tail (Sudhof, 1985). A series of deletion experiments have shown that the seven cysteine-rich repeats are directly involved in apolipoprotein binding (Russell, 1989). Engineered receptors lacking the fifth repeat (amino acids 174-210) display dramatically reduced apo-E binding, suggesting that the receptor interaction with apo-E is largely mediated by this domain (Russell, 1989). The fact that deletion of this domain (and of any

of the other cysteine-rich repeats) does not alter the expression or proper folding of the receptor suggests that the repeats are independently stable and do not rely on significant contacts or disulfide bridges outside of the repeat to be properly folded. Attempts to overexpress the extracellular portion of the receptor have been unsuccessful, thus making it impossible to crystallographically determine the structure of the ligand-binding domain (J. Diesenhofer, personal communication). In light of this, we hope to use a synthetic peptide corresponding to the fifth cysteine-rich repeat as a structural model for the apolipoprotein-binding domain of the LDL receptor.

The protein folding problem (predicting tertiary structure from primary sequence information) is generally recognized as one of the major unsolved computational problems facing modern biology. In the general case, there is no accurate way to predict the structure of a protein unless it is evolutionarily related to a protein whose structure is already known. The LDL receptor cysteine-rich repeat is somewhat unusual, however, in that it is quite small, containing only ≈40 amino acids, and it is most likely stabilized by three internal disulfide bridges. Using a Monte Carlo dynamics procedure that incorporates simulated annealing, recent attempts to predict the structure of crambin (a small protein that also contains three disulfide bridges) were remarkably successful (Wilson, 1989). By applying the same computational procedure to the LDL receptor peptide sequence, we hoped that it would be possible to predict aspects of the folded structure. Results of these simulations, leading to a predicted tertiary structure for the receptor peptide are now described. Preliminary experimental work with the synthetically-made and refolded peptide support some of the basic predictions of the model. The complete three-dimensional structure of the peptide is likely to be determined by x-ray crystallography or NMR in the next two years, allowing a direct assessment of the accuracy of the protein folding algorithm.

Methods

Protein folding algorithm

All attempts to predict the structure of the receptor peptide used the PROSA (PRotein / Optimization by Simulated Annealing) program, available upon request from CW (Wilson, 1989). Details of the method have been published previously but are briefly summarized here for clarity. The algorithm relies upon a simplified model of the protein chain and uses Monte Carlo dynamics to minimize the total energy of the protein. The energy for an arbitrary conformation is evaluated with an empirical energy function that is sensitive to the distance between pairs of amino acids. The protein is represented using a full atom backbone model with ideal internal geometry. The torsion angle about all peptide bonds is fixed in the trans conformation (180°). The backbone torsion angles, ϕ and ψ, are chosen from a probability-weighted Ramachandran plot, derived for each amino acid using the (ϕ,ψ) pairs observed in the data base of known structures. Side chains are represented by a single atom lying at the centroid of the true side chain. Empirical distance-dependent potentials for all amino acid pairings were derived by analyzing the distribution of amino acid pairs found in over one hundred protein structures in the Protein Databank (PDB). Assuming a Boltzmann relationship, these observed probability distributions were converted into distance-dependent energy functions, normalized assuming zero interaction energy for pairs of amino acids separated by 15 Å. Starting from a random conformation, Monte Carlo dynamics is carried out by randomly picking an amino acid in the protein and replacing its current conformation (backbone ϕ,ψ pair) with a new conformation picked from the appropriate Ramachandran plot. The corresponding change in energy is used to determine whether the new conformation is accepted or rejected. Simulated annealing is employed to efficiently overcome local minima and search for the global energy minimum. Tests with crambin indicated that the final structures produced by the folding procedure had

an average r.m.s. deviation relative to the crystal structure of < 5 Å (distance matrix error), but that some structures were as close as 3.2 Å to the true structure (Wilson, 1989).

Receptor peptide sequence

The sequence of the receptor peptide, corresponding to amino acids 169-214 in the human LDL receptor sequence, was obtained from the published data of Sudhof *et al.* (1985). The standard repeat, defined by Russel *et al.* using the multiple copy alignment (Russell, 1989), spans residues 172-210. Amino acids 169-171 and 211-214 have also been included in the folding simulations to take into account the interactions of the repeat with the remainder of the receptor. Alignment of the seven cysteine-rich repeats in the apolipoprotein-binding domain shows that some positions are highly conserved between repeats whereas others vary considerably. Knowledge of the positions of conserved residues was used in the final stages of modelling since invariant amino acids tend to correspond to buried residues.

Folding simulations

A series of computer folding experiments were carried out, initially using only the primary sequence information but gradually including additional constraints developed from previous experiments. At each stage 20-50 structures were generated and average properties for the structures were calculated. The general strategy of the experiments is outlined in Figure 6.1. Initially, a secondary structure prediction was made using the results from the folding simulations in the absence of (ϕ,ψ) restraints. Using the secondary structure prediction as a constraint, it was possible to predict the pattern of disulfide bridges in the folded protein. With both secondary structure and disulfides predicted, a consistent tertiary fold of the peptide resulted from repeated folding simulations. Using manual computer modelling, the position of loop residues was finally adjusted to take into account

knowledge of the conserved residue positions. The conformation of side chain atoms was predicted using the SDCONF program, applied previously to predict side chain positions when modelling protein structure by homology (Wilson, 1991b). Results for each of the simulations are summarized below.

Results

Secondary structure prediction

The secondary structure of the peptide was predicted by initially simulating the folding in the absence of any constraints on (ϕ,ψ) choice and then analyzing the resultant structures for helix and β-strand formation (searching for contiguous blocks of residues whose backbone conformation all correspond to the same type of regular structure) (Wilson, 1989). Figure 6.2 shows the resulting secondary structure profile. A very weak beta-strand signal is found at both the N- and C-termini of the peptide, corresponding to the linker regions between cysteine-rich repeats. Two alpha-helices appear to form the core of the peptide. Both the N-terminal helix (residues 178-185) and C-terminal helix (200-209) are flanked by cysteines. An apparently disordered loop with some weak helical tendencies separates the pair of helices and contains cysteines 188 and 195. In the second series of tests, the observed secondary structure preferences were enforced by restricting the choice of phi-psi pair for those residues with defined helical structure to the appropriate region of the Ramachandran plot (regions defined as described previously (Wilson, 1989)).

Disulfide formation

The receptor peptide contains six cysteines (residues 176, 183, 188, 195, 201, and 210) and there are thus fifteen different combinations of three-disulfide proteins (**1:** 176-183,188-195,201-210; **2:** 176-183,188-201,195-210; **3:** 176-183,188-210,195-210; **4:** 176-188,183-195,201-210; **5:** 176-188,183-201,195-210; **6:** 176-188,183-210,195-201;

7: 176-195,183-188,201-210; **8:** 176-195,183-201,188-210; **9:** 176-195,183-210,188-201; **10:** 176-201,183-188,195-210; **11:** 176-201,183-195,188-210; **12:** 176-201,183-210,188-195; **13:** 176-210,183-188,195-201; **14:** 176-210,183-195,188-201; **15:** 176-210,183-201,188-195). To determine the pattern of disulfides in the folded peptide, all possible three-disulfide combinations were tested as follows. In a given simulation, three distance constraints were introduced to favor the formation of a particular set of three disulfides. After repeatedly simulating the folding of the protein using these constraints, the fraction of final folded structures that had the proper set of disulfides formed was recorded. The formation of a disulfide between a pair of cysteines can be favored by replacing the standard Cys-Cys empirical potential with an artificial potential that includes a deep energy well favoring close side chain contact. Once the contact develops, this favorable interaction is locked in since any change in conformation that tends to separate the pair of cysteines will increase the energy of the protein and be rejected by the Monte Carlo procedure.

Table 6.1 shows the statistics for disulfide formation for the different combinations of distance constraints. Strikingly, only one combination of disulfides appears compatible with the folded structure: 176-210, 183-188, 195-201. All other combinations of disulfide constraints failed to produce a single structure out of fifty attempts in which all three disulfides simultaneously formed. In contrast, the complete set of [176-210, 183-188, 195-201] disulfides formed for a significant fraction of the folded structures produced with these constraints. The reasons for this bias can be partially understood in terms of the helical constraints. Both predicted helices are flanked by cysteines at their N- and C-termini (residues 176, 183, 201, 210). Because the intervening residues are forced into a helical conformation, residues 176 and 183 are always more than 8 Å apart while residues 201 and 210 are always at least 10 Å apart. Thus the combinations that include either the 176-183 disulfide or the 201-210 disulfide (**1,2,3,4,7**) are always prevented from

forming. In addition, when residues 201 and 210 are constrained to a helical conformation, they lie on opposite faces of the same helix. It is thus impossible to dock the second helix parallel or antiparallel to the first and form bridging disulfides across the helix ends (*i.e.* preventing the simultaneous formation of the [176-210, 183-201] disulfides or the [176-201, 183-210] disulfides). This constraint blocks formation of disulfide combinations **12** and **15**. The predicted disulfide pattern differs from that observed in other small cysteine-rich repeats found in receptors (*e.g.* EGF, TGF-α). In surveying disulfide bridging patterns in known structures, Thornton notes that local disulfides (*i.e.* those with a short connecting peptide between the pair of cysteines) are preferred but that often cysteines near the N- and the C-termini are linked to one another (Thornton, 1979). The predicted disulfide pattern thus conforms well to these trends.

Tertiary structure

Given the helical and disulfide constraints, we proceeded to generate a large family of folded structures using the PROSA program. Figure 6.3 shows a superposition of the C_α-tracing for a set of proteins that formed with the presumed native pattern of disulfides. In all structures, the pair of predicted helices cross each other at approximately 60°, separated by 10-15 Å. The lowest energy structure (in terms of the empirical C_α-C_α potential function) is also the structure that has the helices most antiparallel to one another (helical axes crossing at 52°) and separated by the shortest distance (10.2 Å). The helix pair in this structure (shown in bold in Figure 6.3) is similiar to the pair observed in crambin (interhelical angle = 41°, interhelical distance = 9.2 Å), one of the few proteins in which a pair of adjacent alpha-helices are joined by a disulfide bridge. This structure has been used as a basis for further modelling attempts.

Amino acids 169-175 and 211-214 lie outside the core of the repeat defined by the disulfides and thus serve as linkers to adjacent cysteine-rich repeats. The four C-terminal

residues (211-214) usually lie in an extended conformation in the predicted structures, protruding away from the remainder of the protein (Figure 6.3). The N-terminal extension, on the other hand, often folds back onto the peptide and in most structures, the side chain of tyrosine-169 is inserted into the hydrophobic core of the protein. This observation is surprising since residues 164-171 make up an extended linker between repeats four and five which does not exist between any other pair of repeats (Sudhof, 1985)— either the hydrophobic packing in the fifth repeat differs from that found in the other repeats, or the prediction is incorrect with regard to the conformation of the N-terminal extension. We also observed that tryptophan-193, which is well conserved in both the LDL receptor and the LDL receptor-related protein (LRP) repeats (Herz, 1988), was highly exposed in most model structures, pointing away from the peptide. This conformation seems unlikely given the hydrophobic nature of the side chain and its high conservation between repeats. In an effort to improve the prediction, we deleted the three N-terminal residues and rebuilt the loop connecting the helices such that the tryptophan side chain was directed into the hydrophobic core, filling the position previously occupied by tyrosine-169. Following manual rebuilding, the structure was subjected to energy minimization (using the X-PLOR program (Brunger, 1987)) to optimize the geometry of backbone atoms. In the process of rebuilding the loop, backbone torsion angles were adjusted but an effort was made to restrict the (ϕ,ψ) pairs to the observed regions of the Ramachandran plot.

To produce a complete atomic model of the receptor peptide, amino acid side chains were constructed using an algorithm that accurately predicts side chain conformation when modelling protein structure by homology (Wilson, 1991b). This algorithm relies upon a library of side chain rotamers to sample conformation space and evaluates all combinations of rotamers within a local cluster using a free energy force field that includes solvation terms. In previous tests in which side chains were deleted from a known structure and subsequently added using the algorithm, the r.m.s. deviation of side chain atoms from the

original crystal structure coordinates was ≈1.4 Å. This accuracy was not severly reduced in homology-modelling tests, suggesting that the addition of errors in the backbone coordinates should not significantly worsen the prediction.

Discussion

Figure 6.4 shows a stereographic view of the final model of the LDL receptor peptide. The protein consists of two helices, lying at approximately 50° to one another, connected by a single disulfide at the N-terminus of helix one and the C-terminus of helix two. The loop connecting the helices is in an extended conformation, stabilized at each end by a disulfide bridge. The hydrophobic core of the protein contains amino acid side chains from both helices and a single tryptophan from the extended loop. The N- and C-terminal residues lying outside the disulfide-bridged core of the protein are in an undefined conformation in the model, consistent with a role as simple linkers between adjacent cysteine-rich repeats.

Helix-loop-helix structures are found in many different proteins, including the DNA-binding helix-turn-helix motif of many repressor structures and the E-F hand motif of calcium binding proteins. The cysteine-rich repeats of the LDL receptor are known to bind calcium with high affinity and calcium is required for apolipoprotein binding. We compared the predicted structure for the receptor peptide to known E-F hands found in other proteins to determine whether a similiar mechanism for LDL receptor calcium binding is likely. The E-F hand structure is characterized by a pair of helices that meet at a "T", separated by a negatively-charged loop of ≈10 residues that chelates the calcium ion (Richardson, 1988). In both the carp parvalbumin and turkey troponin C crystal structures (Moews, 1975; Herzberg, 1988), at least four acidic amino acid side chains coordinate the divalent cation. In contrast, the constraining 176-210 disulfide forces the helices in the

predicted LDL receptor structure to meet at a "V" (*i.e.* the second helix packs against the first at its end rather than in the middle), and the loop joining the helices is significantly longer than that found in E-F hands. While some sequence features of the predicted receptor loop match those of the E-F hand loop (*e.g.* an abundance of asparates, glycines, and prolines), it does not contain a complete set of four acidic residues to fully coordinate the calcium ion. It is possible that calcium binding sites are formed by the loop without a complete set of carboxylates, or that glutamates and asparates elsewhere on the peptide account for calcium binding.

Given the predicted structure, we can begin to address the following questions: 1) what is the likely accuracy of the model?, and 2) what does the model suggest in terms of the interaction of the receptor with apolipoprotein ligands? To answer the first question, we have analyzed the predicted structure using a number of tests to determine whether it has the gross structural properties expected for a globular protein. By several criteria used to evaluate x-ray refined structures (bond length deviations, bond angle deviations, etc.), the model has good stereochemistry. Backbone torsion angles for all non-glycine residues except cysteine-188 fall well within the allowed regions of the Ramachandran plot. All heteroatoms capable of hydrogen-bonding are predicted to either form an intramolecular hydrogen bond or to be solvent accessible and thus capable of hydrogen-bonding to water.

While these results are encouraging, work by Novonty *et al.* has shown that completely incorrect models of proteins can be refined by energy minimization to produce structures with reasonably good internal geometry (Novotny, 1988). One of the crucial tests of the accuracy of a model-built structure appears to be the extent to which hydrophobic and hydrophilic amino acids are properly partitioned between the buried core and the solvent-accessible surface. Hydrophobic and hydrophilic amino acids are color-coded for identification in Figure 6.4 — almost all polar atoms (excluding those in peptide bonds) are restricted to the surface. In general, the solvent accessibility of each amino acid

correlates with its hydrophocity (data not shown). Hydrophobic amino acids making up the core of the receptor peptide (Phe-179, Leu-184, Ile-189, and Trp-193) are not completely buried. While this observation may suggest problems with the prediction, it is also possible that a small amount of hydrophobic surface is present in the peptide to allow the docking of adjacent repeats within the receptor.

Experimental work to refold and characterize a synthetic peptide corresponding to the fifth cysteine-rich repeat (amino acids 169-214) has been started recently by Elisabeth Jaffe and Mark Wardell (UCSF / Gladstone Labs). Refolding has been accomplished by a two-step procedure in which dialysis of the guanidinium chloride-denatured protein into Tris buffer preceeds disulfide formation and shuffling in a redox-balanced glutathione buffer. Analysis of the refolded peptide by hydrophobic interaction chromatography and molecular sizing chromatography has demonstrated that most of the material exists as a single homogeneous monomeric species. Disulfide mapping of the purified protein (via N-terminal sequencing of isolated proteolytic fragments) argues unambiguously for the formation of the [176-210, 195-201] disulfides. From the current data, it is impossible to conclude if the remaining cysteines [183-188] form a disulfide (although the fact that the peptide is monomeric argues in favor of it). The predicted model, containing the [176-210, 183-188, 195-201] disulfides, thus agrees completely with the preliminary experimental data. Future spectroscopic characterization (*e.g.* circular dichroism, UV fluorescence) of the synthetic receptor peptide should be able to test other aspects of the prediction in advance of the complete structure determination by NMR or x-ray crystallography. The model structure would predict a significant (≈40%) α-helical signal present in the CD spectrum and partial (incomplete) quenching of the tryptophan fluorescence in the folded peptide.

The predicted peptide structure suggests an obvious mechanism by which apolipoprotein-E may recognize the LDL receptor. The second helix is unusually rich in

acidic amino acids, including Asp-200, Asp-203, Asp-206, Glu-207, and Glu-208. For the most part, these amino acids are not paired with basic amino acids and are freely solvent accessible. The distinguishing feature in the electrostatic potential map calculated from the model (Figure 6.5) is a large region of negative potential surrounding this helix. Manually docking the predicted receptor peptide with the LDL receptor-binding domain of apo-E, a series of salt bridges between helix two of the receptor peptide and the basic amino acids in the receptor-binding helix of apo-E can be proposed (Figure 6.6).

Asp-206 is predicted to lie at the center of the second helix in a triplet of amino acids that is rigorously conserved between repeats. The relatively conservative substitution of this amino acid by asparagine is known to virtually abolish apo-E-mediated lipoprotein binding (Esser, 1988). The energetic contribution of a single receptor-ligand salt bridge is unlikely to account for this dramatic effect. However, if the mutation results in the net desolvation of a basic amino acid on the ligand, it could reduce binding dramatically. Accurate modelling of the receptor-ligand complex may be possible by mutating the acidic residues in helix two of the receptor to neutral amino acids and searching for compensatory mutations in the basic region of apo-E. In a series of experiments such as this, it should be possible to map out the energetically-important interactions that account for high affinity binding.

Table 6.1: Disulfide formation in the receptor peptide.

Disulfide constraints	# formed	# complete	<$R_{gyration}$> (Å)	<# contacts>
[176-183, 188-195, 201-210]	46	0	9.270	800.0
[176-183, 188-201, 195-210]	26	0	9.651	782.5
[176-183, 188-210, 195-201]	45	0	9.596	786.0
[176-188, 183-195, 201-210]	49	0	9.265	825.1
[176-188, 183-201, 195-210]	36	0	10.024	761.9
[176-188, 183-210, 195-201]	49	0	9.521	797.8
[176-195, 183-188, 201-210]	77	0	9.301	806.4
[176-195, 183-201, 188-210]	59	0	9.818	786.8
[176-195, 183-210, 188-201]	52	0	9.615	796.4
[176-201, 183-188, 195-210]	89	0	10.120	761.2
[176-201, 183-195, 188-210]	74	0	9.920	789.2
[176-201, 183-210, 188-195]	82	0	9.883	776.8
[176-210, 183-188, 195-201]	**101**	**6**	**9.534**	**789.1**
[176-210, 183-195, 188-201]	77	0	9.521	805.7
[176-210, 183-201, 188-195]	88	0	9.983	776.7

Disulfide constraints were tested as described in the text. Fifty structures were generated with each constraint condition. # formed: total number of disulfides in all structures that formed corresponding to the constraints. # complete: number of structures that had the complete set of disulfides formed. <$R_{gyration}$>: average radius of gyration of the final structures. <# contacts>: average number of long range contacts per structure. Row in bold was identified as the most likely disulfide combination on the basis of the number of structures with the complete set of disulfides formed.

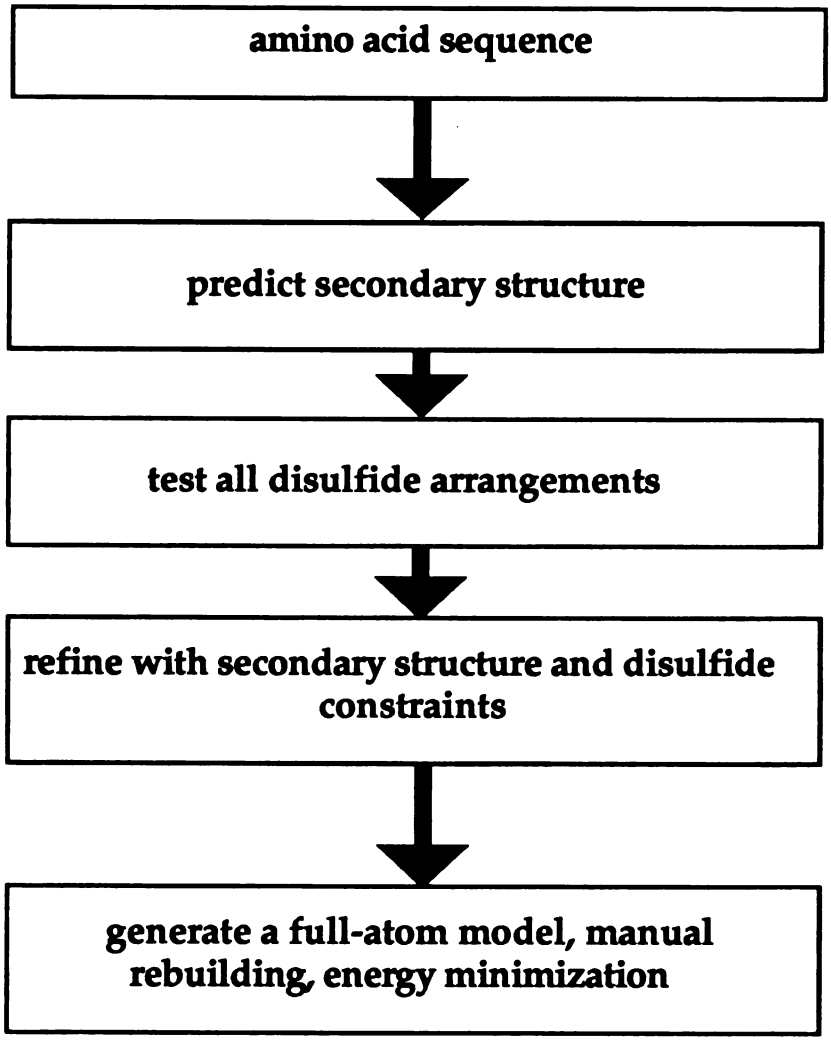


FIGURE 6.1. Prediction scheme for the receptor peptide.

FIGURE 6.2. Predicted secondary structure. Fraction of structures with secondary structure formed at each position is plotted as a function of sequence.

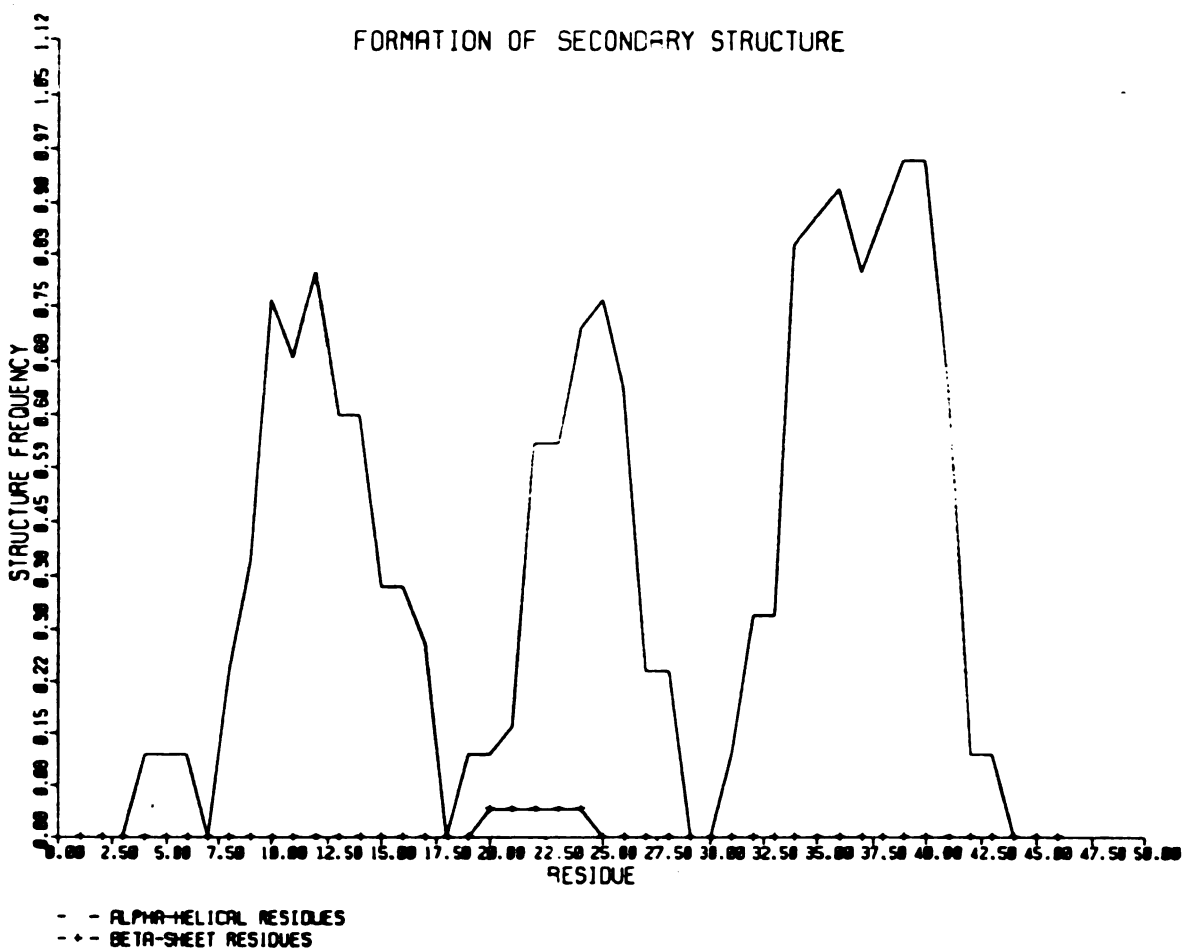


FIGURE 6.3. Superposition of predicted structures. C_α-tracing of six structures produced by the PROSA program using the known secondary structure and disulfide constraints. The lowest energy backbone structure is shown in bold.

FIGURE 6.4. Final model for the receptor peptide. Stereo-view of the fully refined model with hydrophobic amino acids (isoleucine, leucine, valine, phenylalanine, and tryptophan) colored blue and charged amino acids (arginine, lysine, histidine, aspartate, and glutamate) colored yellow.

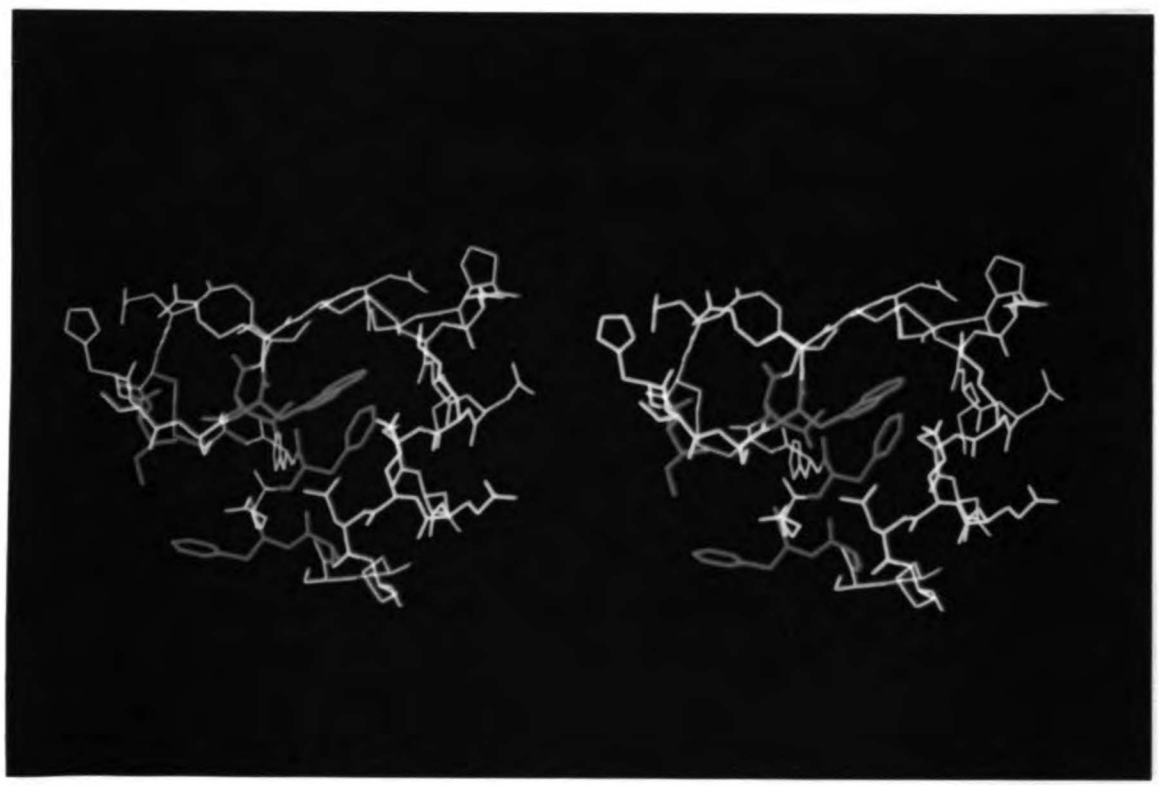


FIGURE 6.5. Electrostatic potential map for the receptor peptide. The DELPHI program (Biosym, San Diego, CA) was used to solve the linearized Poisson-Boltzman equation for the model protein, assuming a solvent dielectric constant of 80 and an ionic strength of 150 mM. +2 and -2 kT/e^- contours are shown in red and blue respectively. Acidic amino acids in helix two (magenta) include Asp-200, Asp-203, Asp-206, Glu-207, and Glu-208.

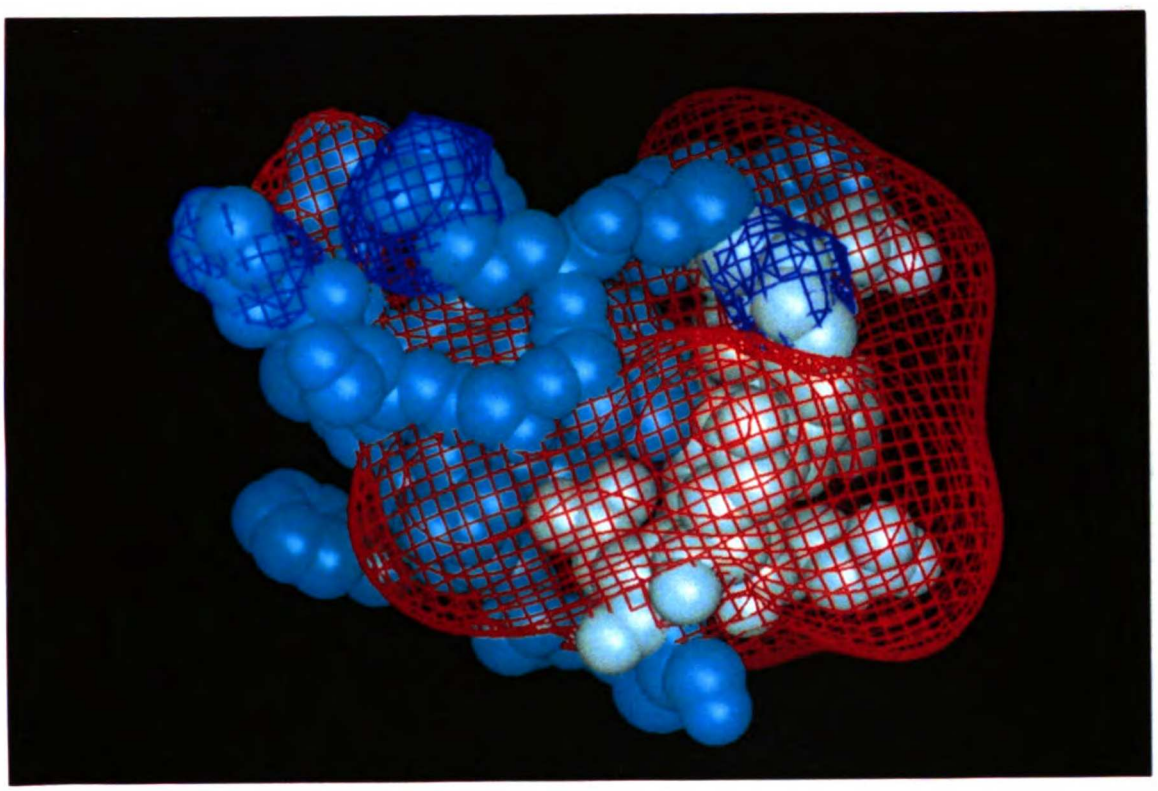
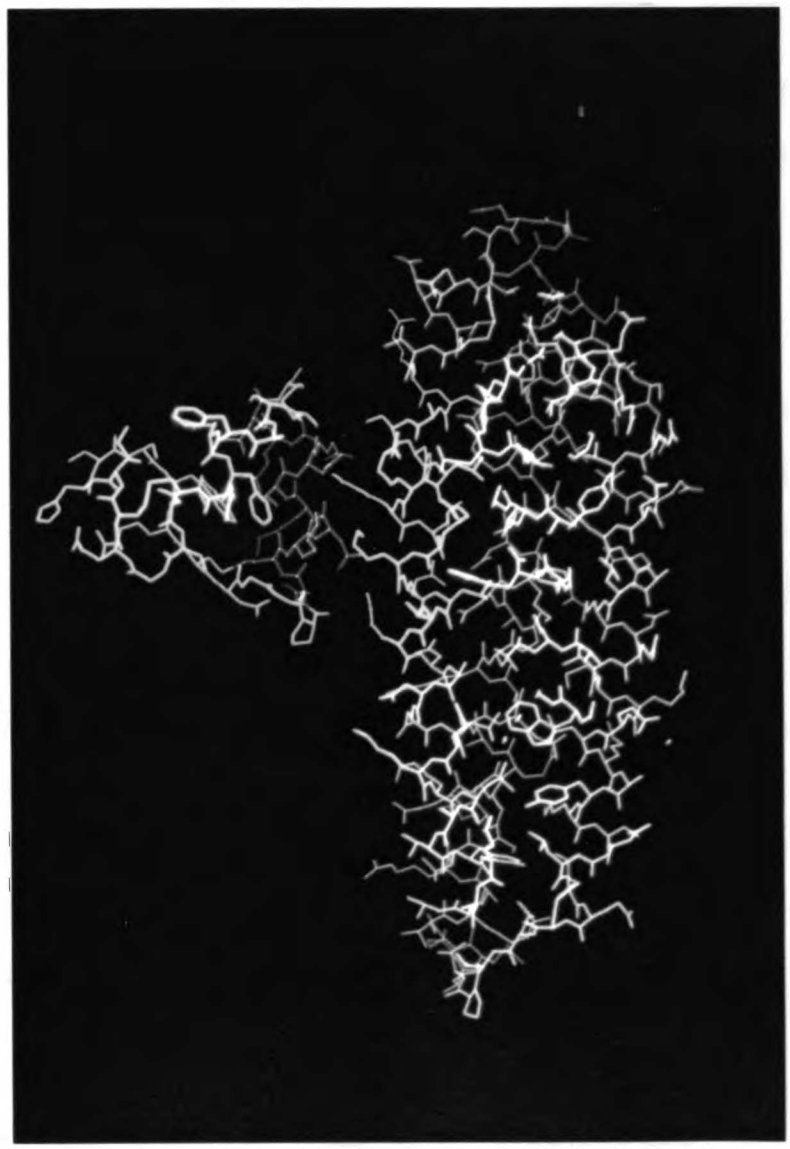


FIGURE 6.6. Predicted docking between the LDL receptor-binding domain of apo-E and the fifth cysteine-rich repeat of the LDL receptor.



References

Brown, M.S., Kovanen, P.T., & Goldstein, J.L. (1980). Evolution of the LDL receptor concept — from cultured cells to intact animals. *Annals New York Academy of Sciences*. 48-68.

Brunger, A.T. (1990) X-PLOR Version 2.1: A system for crystallography and NMR.

Esser, V., Limbird, L.E., Brown, M.S., Goldstein, J.L., & Russell, D.W. (1988). Mutational analysis of the ligand binding domain of the low density lipoprotein receptor. *J. Biol. Chem.* **263**, 13282-13290.

Herz, J., Hamann, U., Rogne, S., Myklebost, O., Gausepohl, H., & Stanley, K.K. (1988). Surface localization and high affinity for calcium of a 500-kD liver membrane protein closely related to the LDL-receptor suggest a physiological role as lipoprotein receptor. *EMBO Journal.* **7**, 4119-4127.

Herzberg, O. & James, M.N.G. (1988). Refined crystal structure of troponin C from turkey skeletal muscle at 2.0 angstroms resolution. *J. Mol. Biol.* **203**, 761.

Moews, P.C. & Kretsinger, R.H. (1975). Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. *J. Mol. Biol.* **91**, 201.

Novotny, J., Rashin, A.A. & Bruccoleri, R.E. (1988). Criteria That Discriminate Between Native Proteins and Incorrectly Folded Models. *Proteins: Struct. Func. Genet.*, **4**, 19-30.

Pitas, R.E., Innerarity, T.L., & Mahley, R.W. (1980). Cell surface receptor binding of phospholipid·protein complexes containing different ratios of receptor-active and -inactive E apoprotein. *J. Biol. Chem.* **255**, 5454-5460.

Richardson, J.S. & Richardson, D.C. (1988). Helix lap-joints as ion-binding sites: DNA-binding motifs and Ca-binding "EF hands" are related by charge and sequence reversal. *Proteins: Structure, Function, and Genetics.* **4**, 229-239.

Russell, D.W., Brown, M.S., & Goldstein, J.L. (1989). Different combinations of cysteine-rich repeats mediate binding of low density lipoprotein receptor to two different proteins. *J. Biol. Chem.* **264**, 21682-21688.

Sudhof, T.C., Goldstein, J.L., Brown, M.S., & Russell, D.W. (1985). The LDL receptor gene: mosiac of exons shared with different proteins. *Science.* **228**, 815-822.

Wilson, C. & Doniach, S. (1989). A computer model for dynamically simulating protein folding: studies with crambin. *Proteins: structure, function, and genetics.* **6**, 193-209.

Wilson, C., Wardell, M., Weisgraber, K.H., Mahley, R.W., & Agard, D.A. (1991a). The three-dimensional structure of the LDL-receptor binding domain of human apolipoprotein-E. *Science.* in the press.

Wilson, C., Gregoret, L.M., & Agard, D.A. (1991b). Modelling side chain conformation for homologous proteins using an energy-based rotamer search. submitted for publication to *J. Mol. Biol.*

Chapter 7 :

A computational method for the design of enzymes with altered substrate specificity.

Abstract

A combination of enzyme kinetics and x-ray crystallographic analysis of site-specific mutants has been used to probe the determinants of substrate specificity for the enzyme α-lytic protease. We now present a generalizable model for understanding the effects of mutagenesis on enzyme substrate specificity. This algorithm uses a library of side chain rotamers to sample conformation space within the binding site for the enzyme-substrate complex. The free energy of each conformation is evaluated with a standard molecular mechanics force field, modified to include a solvation energy term. This rapid energy calculation based on coarse conformation sampling quite accurately predicts the relative catalytic efficiency of over forty different α-lytic protease–substrate combinations. Unlike other computational approaches, with this method it is feasible to evaluate all possible mutations within the binding site. Using this algorithm, we have successfully designed a protease which is both highly active and selective for a non-natural substrate. These encouraging results indicate that it is possible to design altered enzymes solely on the basis of empirical energy calculations.

Introduction

Many biological processes are characterized by the formation of a highly specific macromolecular complex. In an attempt to understand the energetic terms important for molecular recognition, our laboratory has been studying the interaction between the enzyme α-lytic protease and its peptide substrates (Bone *et al.*, 1987; Kettner *et al.*, 1988; Bone *et al.*, 1989a; Bone *et al.*, 1989b; Bone *et al.*, 1990). We have used a wide variety of peptide substrates to obtain accurate estimates for the binding energy of the reaction transition state. This functional data has been complemented by crystallographically determined high resolution structures of enzymes complexed to a parallel set of peptide inhibitors. Site directed mutagenesis of residues in the substrate binding pocket has provided a large number of variant enzymes with altered activity. Kinetic and crystallographic characterization of these mutants has led to a better understanding of the factors important in determining substrate specificity. In the course of these studies, we have developed a data base of 42 enzyme-substrate kinetics measurements and structures and K_i data for 18 different enzyme-inhibitor complexes. This large body of data provides a rigorous test of the accuracy of any algorithm designed to predict the energetic effects of altering a macromolecular complex.

There is considerable interest in engineering biological molecules to form specific complexes (*e.g.* rational drug design, enzyme engineering). Discovering which mutations to a protein or alterations of functional groups on a drug will enable it to form a tight complex is a combinatorial problem in which thousands of candidate molecules may need to be tested. Recently-developed free energy perturbation (FEP) methods should theoretically be able to accurately calculate the effects of mutagenesis on an enzyme-substrate complex or for a drug binding to its target. A number of test cases have been

reported in which the components of a complex are slightly altered and the energetic consequences are successfully predicted by FEP (Warshel *et al.*, 1988; Rao *et al.*, 1987). While these methods can yield accurate energy estimates for some cases, they are extremely computationally intensive and thus have limited utility in the more general problem of drug design or protein engineering.

Here we describe the development and application of an energy-based conformation search method to the problem of predicting changes in the stability of an enzyme-substrate complex. Conformation space is broadly sampled using a library of idealized rotamers. The energy of each rotamer combination is evaluated using a molecular mechanics force field which contains an additional solvation energy term. Remarkably good agreement is found between the calculated substrate binding energies and the experimentally-observed values for α-lytic protease. We have used this algorithm to design a highly active enzyme with a desired alteration in substrate specificity. This method should thus be a generally useful tool for both computer-based drug design and for the design of mutant enzymes.

Methods

Figure 7.1 outlines the various steps to our algorithm. To calculate the relative change in the stability of a macromolecular complex, we individually calculate the free energy for each of the molecules in the complex and then for the complex itself. All allowed conformations are evaluated and the corresponding free energies are used to calculate the overall free energy for each molecule:

$$\Delta G = \Delta G_o - RT \ ln \sum_i exp \left(\frac{-(\Delta G_i - \Delta G_0)}{RT} \right) \qquad (1)$$

where ΔG_o is the ground state energy, ΔG_i is the energy of the *i-th* conformation, and the summation is done over all allowed conformations. The difference between the free energy of the complex, ΔG_{ES}, and the free energy sum of the free substrate and enzyme, (ΔG_E + ΔG_S), yields a relative ΔG for complex formation. Repeating these calculations on a slightly modified complex gives another energy, $\Delta G'$, and the difference between them ($\Delta\Delta G=\Delta G'-\Delta G$) directly determines the energetic cost of the modification.

(a) *Rotamer representation of conformation space*

The challenge in using such an energetic approach is to develop a practical method for adequately sampling conformational space. To limit the set of conformations to a computationally manageable number, we have ignored alternate enzyme and substrate main chain conformations and fixed the backbone atoms using coordinates from a wild type enzyme-inhibitor structure. Our representation of side chain conformation space is based on the recent observations of Ponder and Richards (1987a). By examining a collection of high resolution protein crystal structures, these authors have shown that the conformations sampled by amino acid side chains are relatively limited and can be reasonably approximated as a set of idealized rotamers. These rotamers, specified by the torsion angles between non-hydrogen side chain atoms, generally correspond to staggered low-energy conformations. Because their work is based on a small set of high resolution structures, amino acids defined by many torsion angles (*e.g.* methionine) are relatively under sampled and thus not all possible rotamers are included in the Ponder-Richards library. To correct for this, we examined all staggered side chain conformations and added those with no bad intraresidue contacts to create a new library (thereby introducing another asparagine rotamer and nine additional methionine rotamers). 95 rotamers are used to represent the possible conformations of all amino acids except lysine and arginine. The

complete conformation space of a defined site is given by all of the different rotamer combinations for the side chains within the site. Our calculations for α-lytic protease have used a site which contains the P1 substrate residue and those enzyme residues within 4Å of it (a total of four amino acids).

(b) *Energy calculation*

The free energy for each conformation is calculated as the sum of a non-bonded term, E_{NB}, and a solvation term, ΔG_{solv}:

$$\Delta G = w_{NB}E_{NB} + w_{solv}\Delta G_{solv} \qquad (2)$$

where w_{NB} and w_{solv} are empirically determined weights for the non-bonded and solvation terms respectively. The reference state for the force field (conditions under which $\Delta G = 0$) assumes that 1) the atoms in each molecule are non-interacting with respect to electrostatic, hydrogen-bonding, or van der Waals terms, and 2) the atoms are completely buried in a protein-like hydrophobic environment.

The non-bonded contribution is estimated using our implementation of the AMBER all-atom molecular mechanics force field (Weiner *et al.*, 1984). Since the rotamers calculated from the standard library have idealized internal geometry, the energy for distorting bond lengths and bond angles can be neglected. The interactions of hydrogen atoms are included, using the PROPAK program to calculate their atomic coordinates (Ponder & Richards, 1987b). For hydroxyl protons whose position is undetermined (*e.g.* serine HOG), all staggered rotamers for the proton were evaluated. We have incorporated the following changes to the standard AMBER force field. Rather than using partial charges for every atom type, partial charges were assigned only to atoms with an absolute charge $|q| > 0.2e^-$. All polar hydrogens were assigned the same partial charge ($+0.31e^-$). Instead of using a multipole representation for sulfur atoms (nucleus + two electron lone

pairs), sulfurs were represented as a single neutral atom. Non-bonded interactions were arbitrarily capped at 100 Kcal/mole and were evaluated between 1) all pairs of rotamer atoms within a site, and 2) between pairs of fixed and rotamer atoms for fixed atoms that lay within an 8Å box enclosing the site. A distance-dependent dielectric, $\varepsilon = r$, was used to scale electrostatic interactions.

The solvation term was evaluated using an approach based on the method of Eisenberg and McLachlan (1986). In their formalism, the free energy of transfer for an atom i is calculated as:

$$\Delta G_i = \Delta\sigma_i A_i \qquad (3)$$

where ΔG_i is the free energy change in bringing atom i from a hydrophobic protein interior to aqueous solution, $\Delta\sigma_i$ is an atomic solvation parameter, and A_i is the solvent-accessible surface area of the atom. In the original Eisenberg-McLachlan algorithm, a molecular surface (as defined by Lee and Richards (1971)) must first be calculated to estimate A_i. The calculation of a molecular surface is a computationally intensive process which can not feasibly be done when carrying out a combinatorial rotamer search (in which thousands of different structures are generated for each site). We have developed an alternative grid-based approach for estimating A_i which increases the speed of the solvation calculation by several orders of magnitude. As a first step, positions in a body-centered cubic lattice, spanning the entire site and corresponding to allowed water positions, are calculated. A water molecule is assumed to occupy a lattice point if it is not within van der Waals contact distance of any other protein atoms and if it forms a minimum number of hydrogen bonds. The van der Waals contact radii defined by Shrake and Rupley (1973) are used to define the protein excluded volume. The solvent accessible surface area for each atom is then calculated as:

$$A_i = \frac{4\pi r^2 \cdot n}{62} \qquad (4)$$

where r is the atom's van der Waals contact radius and n is the number of solvent-occupied lattice points lying within a shell 3.3—3.6Å from the atom (a completely exposed atom has 62 neighboring lattice points occupied). Non-bonded interactions between real atoms and the lattice solvent molecules are not considered since these are assumed to be accounted for by the solvation term. For aspartate, glutamate, and arginine side chains, both terminal heteroatoms are assigned the charged atom (N+/O-) atomic solvation parameter (see below).

(c) *Structural model of the enzyme-substrate complex*

Peptide boronic acids form transition state-like complexes with serine proteases and serve as good structural models for the active complex between substrate and enzyme (Matthews *et al.*, 1970; Bone *et al.*, 1987). For the α-lytic protease test case, we used the x-ray crystal structure of the wild-type enzyme complexed to the methoxy-succinyl-alanine-alanine-proline-borovaline peptide inhibitor (PDB name 1P03) as a basis for all calculations. This is the highest affinity wild-type enzyme-tetrapeptide complex for which we have structural information (K_i = 6 nM, Bone *et al.*, 1989b). The conformation of the enzyme and the substrate observed in the complex was used to model the free molecules. All residues were held fixed except for the amino acids in the S1 pocket of the enzyme [192, 213, and 217A, chymotrypsin-numbering scheme (James *et al.*, 1978)] and the P_1 residue of the substrate. To account for the additional flexibility of the substrate relative to the enzyme, the C_β atom of the P_1 side chain was allowed to swing ±5° perpendicular to the plane defined by the N,C_α, and C atoms of the P_1 residue. A complete set of rotamers were then calculated for each C_β position. Different side chain substitutions were made by generating atomic coordinates for all rotamers at each of the varying sites [using the Ponder-Richards PROPAK program (1987b)]. All rotamers corresponding to the

appropriate amino acid at each of the varying sites were then combinatorially evaluated. As an example, Figure 7.2 shows the rotamers used to sample conformation space for a representative enzyme-substrate combination. In this figure all rotamers are shown simultaneously.

(d) *Enzyme-substrate kinetics*

The relative binding energy between the substrate and the enzyme in the transition state can be obtained directly from experimental measurements of k_{cat} and K_M using the following formula (Fersht, 1985):

$$\Delta G^{\dagger}_T = -RT\ ln\ (\frac{k_{cat}}{K_M}) - RT\ ln(\frac{kT}{h}) \qquad (5)$$

Table 7.1 summarizes the kinetics measurements we have determined for α-lytic protease with substrates of the type succinyl-alanyl-alanyl-prolyl-X-*p*-nitroanilide (where X = all twenty naturally occurring amino acids except for proline, tryptophan, arginine, and lysine). It is important to note that the experimental data contains information from polar, non-polar, and charged residues and thus provides a stringent test for all aspects of the energetic model. The following proteases have been included in the data set: wild type, Met→Ala192, Met→Ser192, Met→Ala213, and Val→Ala217A. Each $\Delta\Delta G$ is calculated relative to the wild type enzyme for the alanine substrate, using the formula:

$$\Delta\Delta G(i) = -RT\ ln\ (\frac{k_{cat}}{K_M})_i + RT\ ln\ (\frac{k_{cat}}{K_M})_{WT+ala} \qquad (6)$$

This arbitrary reference state was chosen because alanine is the best substrate for the native enzyme and has been used as the starting point for other calculations of enzyme-substrate specificity for α-lytic protease (Caldwell *et al.*, 1990).

(e) *Parameterization*

The data set of kinetics measurements was randomly segregated into two groups. Twenty measurements were used to determine the scale factors in equation 2 which optimize the correlation between calculated and experimental binding energies (the parameterization data set). Because the calculated binding energies are a non-linear function of the free parameters, the scale factors were iteratively adjusted until the correlation was maximized. These optimized parameters were then used to calculate binding energies for twenty other enzyme-substrate combinations (the testing data set). Table 7.2 indicates which measurements formed each data set.

Results

(a) *Parameterizing the solvation model*

To accurately calculate solvation energies, the atomic solvation parameters (ASPs) for different atom types were adjusted such that the fit to experimental *n*-octanol→water transfer free energies of Fauchere & Pliska (1983) for N-acetyl amino acid amides was optimized. Figure 7.3 shows a plot of the calculated versus experimental values. The correlation coefficient between the experimental and calculated transfer free energies gives a measure of the accuracy of the solvation model. Using five atom types (carbon, neutral nitrogen, neutral oxygen, charged nitrogen/oxygen, and sulfur), both the molecular surface approach of Eisenberg and McLachlan (1986), and the grid approach developed here yield a correlation coefficient of 0.95. This suggests that the errors in A_i introduced by the grid

approach are less significant than the errors inherent to the formalism defined by equation 3. The ASPs for the five atom types listed above were refined to the following values:

Carbon	20±3 cal/mol/Å^2	N(+)/O(-)	-112±13 cal/mol/Å^2
Nitrogen	-40±9 cal/mol/Å^2	Oxygen	-64±11 cal/mol/Å^2
Sulfur	53±20 cal/mol/Å^2		

All of the enzyme-substrate complexes considered in the data set contain at least one methionine in the active site. Given that the sulfur atomic solvation parameter is determined by only two experimental measurements (methionine and cysteine transfer free energies), we questioned the accuracy of this ASP. Eisenberg and McLachlan note that it is impossible to determine a reliable ASP for the three different sulfur atom types (methionine, cysteine, cystine) and in their calculations the cysteine value is ignored. Early comparison of the calculated and experimental binding energies suggested that a better fit could be obtained if the hydrophobicity of the methionine sulfur was increased. The ASP for methionine sulfur atoms was thus increased to 80 cal/mol/Å^2 and this value was used in all subsequent calculations.

(b) *Optimizing the model*

In modelling the effects of mutagenesis on enzyme-substrate specificity, we have assumed that the enzyme-inhibitor complex is an accurate representation of the transition state for substrate hydrolysis. Figure 7.4 shows a plot of the experimental binding energies for boronic acid inhibitors versus the experimental values for the corresponding substrates. In general there is good agreement, indicating that this basic assumption is reasonable. The correlation between the two sets of energies (0.87) sets an upper limit on how well we might expect to calculate substrate binding energies.

In principle, the non-bonded and solvation contributions to the energy are already properly scaled ($w_{NB} = w_{solv} = 1$) and could be used directly to calculate binding energies. Table 7.2 shows the results obtained using this assumption. The correlation between all calculated and experimental binding energies is 0.26 ($P > 0.10$) and the slope of the best-fit line is 0.013. An analysis of the results clearly shows that the van der Waals interactions between atoms in close contact with one another are drastically overestimated for certain complexes. This is not surprising since the rotamers used to model the active site are fixed and thus unable to accommodate slight bad bumps. The electrostatic terms also appear to be overweighted, giving unreasonable results. For example, the calculated difference between serine and alanine substrates binding to the Met192→Ser mutant (-23.5 Kcal/mole) is an order of magnitude greater than the experimentally observed difference (+1.2 Kcal/mole). The calculated difference in binding energies drops to -0.9 Kcal/mole if electrostatic terms are ignored.

The above observations suggest that the various components of the force field are not properly scaled to one another and should be adjusted to take into account the various assumptions of the model, *e.g.* that side chains are rigid idealized rotamers, that the protein backbone is fixed, etc.. Because the ASPs were parameterized using experimental data of Fauchere and Pliska (1983), the solvation model assumes that the hydrophobicity in the buried protein is mimicked by *n*-octanol. Previous studies have indicated that proteins are somewhat more hydrophobic than *n*-octanol, (Dorovska-Taran *et al.*, 1982), indicating that the solvation term may need to be upweighted. The relative weights of the non-bonded component and the solvation component to the total energy were therefore adjusted to maximize the correlation between calculated and experimental binding energies for the parameterization set of data points. Scaling the non-bonded term by 0.031 and the solvation term by 1.98 yielded an overall correlation of 0.77 ($P < 10^{-8}$) and a best-fit line with a slope of 1.06 (results shown in table 7.2).

As shown in equation 1, the free energy calculated for a molecule contains a temperature-dependent term representing the entropic contribution of non-ground state conformations [-RT ln Σ exp(-$\Delta\Delta G$/RT)]. Equation 1 is valid if the summation over discreet states is a good approximation of the integral over all possible states. In sampling a limited number of rotamers, we are poorly estimating the contribution of non-rotamer conformations and may thus underestimate the partition function. In addition, if components to the free energy are not properly scaled, the value of the entropic term will be distorted because the partition function is a non-linear combination of the free energies for each conformation. For these combined reasons, it seemed likely that additional parameters could help offset errors in the model and allow a more accurate calculation of the contribution of non-ground state conformations to the free energy. These parameters were introduced by scaling the $\Delta\Delta G$s before calculating the summation in equation 1 and by applying a compensatory overall scale to properly weight the entropic term. The two parameters had strongly coupled effects on the correlation coefficient and hence are not truly independent. After optimization [during which the overall scale factor rose to 5.7 and the $\Delta\Delta G$ scale factor inside the summation rose to 25], the correlation between calculated and experimental binding energies increased to 0.88. The significant improvement in the correlation validates use of the additional parameters as judged by the χ^2-test. The non-physical nature of these additional scale factors makes their validity for other applications questionable, although their introduction definitely improves the ability to reproduce experimental data for the α-lytic protease test system.

An important validation of the true capabilities of this approach are to test the model on data not used in the parameterization. Using optimized parameters derived from a subset of twenty kinetics values, the correlation coefficient for the remaining twenty observations was 0.85 ($P < 10^{-11}$, data shown in table 7.2). Performance on the complete data set of 40 measurements also gave a 0.85 correlation coefficient. We repeatedly re-

segregated the data points into new parameterization and testing sets and re-optimized to determine how sensitive the parameters were to the particular choice of data. The final parameters generally fell within 5% of those reported above and correlation coefficients averaged 0.85.

(c) *Importance of different aspects of the model in reproducing experimental data*

Given a statistically significant correlation between calculated and experimental values for the binding energies, we can ask what aspects of the algorithm are essential for accurately reproducing experimental results. Since parameters have been adjusted to optimize the fit to experimental values, any changes in the force field would be expected to worsen the correlation. The importance of each part of the algorithm can be assessed by removing it from the calculation and then noting the drop in the correlation obtained using the recalculated binding energies. These results are summarized in table 7.3.

We have assumed that conformation space must be broadly sampled (using a complete search of all rotamer combinations) to accurately estimate the free energy change for a particular mutation. To test this assumption, we have repeated the calculation using a single rotamer at each site. When only the most commonly observed rotamer in the rotamer library is used for each amino acid, the correlation between calculated and experimental binding energies drops from 0.85 to -0.26. Carrying out the rotamer search is thus a key part of the algorithm. The temperature dependence of the correlation indicates how the non-ground state rotamer combinations contribute to the overall $\Delta\Delta G$. If the temperature is arbitrarily set to zero, the correlation drops to 0.75. Thus while the ground state (lowest energy) rotamer combination provides most of the information needed to estimate the energy of a complex, the entropy term given by the ratio of the partition functions is also informative and improves the ability to reproduce experiment.

By omitting the non-bonded or solvation terms from the energy calculation, we can estimate how important either is in fitting experimental values. When only non-bonded terms are included, the correlation between calculated and experimental binding energies drops from 0.85 to -0.09. Using only solvation terms, the correlation drops to 0.36. Clearly both terms are essential for obtaining a good fit. It is important to note that the non-bonded terms contribute significantly to the correlation, despite the relatively low weight given to them after optimization ($w_{NB} = 0.031$).

The ΔG calculated for complex formation is a function of the energy of the free enzyme, the free substrate, and of the complex itself. If either the ΔG_S or ΔG_{ES} term is ignored, the correlation drops dramatically (data shown in table 7.3). If the ΔG_E term is neglected from the calculation, *i.e.* if the free energy for the unbound enzyme is assumed to be constant, the correlation drops significantly to 0.66.

(d) *Design of a protease with altered substrate specificity*

To rigorously test the algorithm, we have attempted to design a mutant protease which is both specific and highly active for a substrate not normally hydrolyzed by the wild type enzyme. All possible single-site mutations within the binding pocket were computationally evaluated for the ability to cleave substrates with P_1 = leucine, since leucine is a poor substrate for the native enzyme. An enzyme active for leucine substrates might also be expected to cleave isoleucine substrates since these residues are both structurally and chemically similar to one another. In fact, the difference in binding energy for the two substrates, $\Delta\Delta G_{Leu}$-$\Delta\Delta G_{Ile}$, for the wild type enzyme was calculated to be only -0.16 Kcal/mole (prior to this work, there were no experimental kinetics data for the wild type enzyme with isoleucine substrates). We thus screened all possible mutants for those predicted to prefer leucine over isoleucine substrates by at least 3 Kcal/mole, corresponding

to a factor of ≈150 in relative activity. From this subset of mutants, the substitution Met192→Val was predicted to yield the enzyme most active for leucine. The predicted binding energy for the leucine substrate ($\Delta\Delta G$ = 0.85 Kcal/mol, relative to the Ala substrate), is significantly better than that for the isoleucine substrate ($\Delta\Delta G$ = 7.90 Kcal/mol).

The Met192→Val mutant was made, expressed, and purified using techniques described previously (Bone *et al.*, 1989a). Figure 7.6 and table 7.4 show the kinetic data obtained for the wild type and mutant enzymes with either leucine or isoleucine substrates. As predicted, wild type α-lytic protease slightly prefers leucine over isoleucine substrates ($\Delta\Delta G$ = -0.42 Kcal/mole) but cleaves both poorly. The k_{cat}/K_M value for the designed Met192→Val mutant with the leucine substrate has been dramatically increased to 5320 $s^{-1}M^{-1}$, corresponding to a binding energy of 0.82 Kcal/mole relative to the reference state (k_{cat}/K_M = 21,000 $s^{-1}M^{-1}$ for the wild type enzyme with the alanine substrate). This free energy difference is essentially identical to the $\Delta\Delta G$=0.85 Kcal/mole predicted by the algorithm. The mutant shows significantly less activity towards the isoleucine substrate (k_{cat}/K_M = 24.0 $s^{-1}M^{-1}$, $\Delta\Delta G(exp)$ = 4.01 Kcal/mole).

While the enzyme is quite selective for leucine over isoleucine (preferring leucine >200-fold), the isoleucine substrate is significantly more active than predicted. We examined the ground state rotamer structure for the mutant complex with isoleucine to understand why the calculated binding energy is so unfavorable. The structure predicts several bad van der Waals contacts between the C_δ methyl group of the P_1 side chain and other atoms in the active site. It appears that these bad contacts could be partially relieved by a 45° rotation of the methyl group. Because our conformation search is limited to idealized rotamers, we do not consider this possibility in evaluating the isoleucine substrate. In addition, the model is based on a rigid backbone which in reality would be

expected to relax to minimize the bad contacts. These effects could explain why this algorithm overestimates $\Delta\Delta G$ for bad substrates such as isoleucine.

Discussion

We have applied an energy-based conformation search to estimating the energetic effects of mutagenesis on the formation of a macromolecular complex. To estimate the relative free energy, the Ponder and Richards (1987a) side chain rotamer library has be used to sample conformation space, and a combination of an AMBER molecular mechanics force field (Weiner *et al.*, 1984) and an additional ASP-based solvation term (Eisenberg & McLachlan, 1986) have been used to evaluate the energy for each conformation. This algorithm successfully models the results of a large body of experimental data for α-lytic protease. In addition, the method has been used to design a mutant protease with high activity and high selectivity for a non-natural substrate (leucine).

Previous attempts at engineering the components of a macromolecular complex have been limited by both the accuracy and the speed of existing computational methods for calculating free energy changes. Empirical, heuristic methods for estimating binding energies between macromolecular complexes have been described by Naray-Szabo (1989) and by Novotny *et al.* (1989). The most widely cited approach, based on free energy perturbation (FEP), can be used to obtain highly accurate estimates of the $\Delta\Delta G$ for a well-defined mutation in a complex (Warshel *et al.*, 1988; Rao *et al.*, 1987). For example, Bash *et al.* (1987) have shown that FEP can successfully predict the effect of replacing the amide group of a phosphonamidate inhibitor of thermolysin with an ester (substituting -NH- for -O-)[19]. For reasons discussed below, however, FEP has somewhat limited utility in the more general problem of designing mutations which will form a tight complex.

To estimate energy changes using FEP, a molecular dynamics simulation is carried out in which the Hamiltonian describing the system is gradually perturbed from the native state to the mutant state. If the starting and final states are structurally different, the simulation should be long enough to allow the conformational transition between the two states to occur (Bash *et al.*, 1987). Since the rotation of medium-sized side chains buried in a protein has a characteristic time of 10^8-10^{12} picoseconds (McCammon & Harvey, 1987), it is impossible to accurately simulate this type of transition by molecular dynamics. This suggests that if mutagenesis significantly alters the structure of a well-packed complex (*e.g.* if a side chain rotation occurs), FEP will be unable to predict the energetic consequences of the change. Crystallographic studies comparing peptide boronic acids bound to different α-lytic protease mutants have shown that a mutation at one site can lead to significant alteration in the surrounding side chain conformations (Bone *et al.*, 1989a). These changes can have dramatic effects on substrate specificity.

A second problem facing free energy perturbation methods is the amount of computer time needed to simulate a single mutation. Because the molecular dynamics simulation must be carried out near equilibrium to obtain accurate results, each $\Delta\Delta G$ calculation typically requires hours of supercomputer time. In the case of enzyme design or drug design, thousands of different mutations are potentially worth evaluating. The inability to rapidly screen through a large number of different complexes may explain why FEP methods have not yet succeeded in the rational design of novel pharmaceuticals or altered enzymes.

Recent calculations on the substrate preferences for wild type α-lytic protease (Caldwell *et al.*, 1990) allow a direct comparison of the accuracy of free energy perturbation methods and of the algorithm described here. Caldwell *et al.* have calculated the transition state binding energies for glycine, valine, and leucine substrates, relative to the alanine substrate, using free energy perturbation. Their estimates for the glycine and

valine relative binding energies differ from the experimental values by -0.72 Kcal/mole and +2.27 Kcal/mole respectively. Our estimates for these binding differences are somewhat more accurate, with an error of -0.08 Kcal/mole for glycine and -1.80 Kcal/mole for valine. The free energy perturbation calculation drastically overestimates the difference in binding of leucine versus alanine substrates ($\Delta G(calc) > 20$Kcal/mole), whereas our calculation underestimates the energy difference by 1.38 Kcal/mole. Crystallographic analysis using a peptide-boroLeu inhibitor, revealed that although structural changes occur upon leucine binding, the dominant change is simply the rotation of the side chain of Val217A by 120°. This result underscores the practical difficulties in obtaining reliable values by the FEP method when conformational alterations are required. While these results clearly favor the algorithm developed here, it is important to note that the method has been parameterized using a subset of the α-lytic protease data. Until the algorithm is applied to other experimental systems, it remains unclear whether it will consistently out perform FEP.

A major goal of this work was to produce an algorithm that was both accurate and fast enough to be useful in the rational design of enzymes with altered specificity. In the example of α-lytic protease, there are at least three residues (192,213, and 217A) which are likely to alter substrate specificity. To test all possible mutations at these sites for a single substrate ($20^3 = 8000$ sequences, or $\approx 111^3 = 1.4 \cdot 10^6$ rotamer combinations), one must be able to automatically screen through each mutant rapidly. The current algorithm calculates a binding energy for each rotamer combination in approximately $2.2 \cdot 10^{-4}$ sec VAX 8650 CPU time. Complete combinatorial mutagenesis of the binding pocket for a given substrate thus requires less than three hours.

A remaining question is whether the scale factors obtained for the α-lytic protease test case are in fact the optimal parameters to use for other systems. All protein backbone atoms are held fixed in our algorithm; in cases in which the active site actually expands to accommodate large substrates, we therefore tend to overestimate the van der Waals

interactions between the substrate and the protein. To compensate for this assumption, non-bonded terms have been significantly down weighted relative to the original force field values. The amount that main chain atoms can relax to accommodate bad contacts is presumably a characteristic of a given site, and thus the plasticity of the α-lytic protease active site may require a non-bonded weight significantly different from that used for other proteins. If the active site of another protein is considerably more rigid than the α-lytic protease active site, for instance, a non-bonded scale factor closer to one may be needed since the assumption that the backbone remains fixed would be more accurate. Since the scale factors used to obtain good agreement for the parameterization data set ($r = 0.88$) also yield good results for the testing data set ($r = 0.85$), these parameters appear to be well determined for the α-lytic protease system. We are in the process of applying the same free energy force field to the problem of predicting side chain conformation. Quite good results are obtained using the same set of scale factors determined for the α-lytic protease data, suggesting that the values used are acceptable for other applications.

The results of this simple rotamer-based conformation search are extremely encouraging — they suggest that it is feasible to design enzymes with desired catalytic specificities solely on the basis of computer calculations. A major advantage of the method described here is its ability to rapidly sample sequence space to select combinations of mutations which may have altered binding properties. Even higher accuracy in predicting the effects of mutagenesis on substrate binding energies might be obtained by combining this coarse conformation searching algorithm with other computational methods that sample local conformations more finely (*e.g.* molecular dynamics, free energy perturbation methods). Future algorithms will attempt to apply a 'high-resolution' optimization of local geometry to those configurations which appear feasible after a 'low-resolution' rotamer search. In addition, by expanding the initial rotamer search to also consider shifts in

backbone atoms, we hope to more broadly sample the full conformational space that is important in determining the effects of mutagenesis on complex formation.

Table 7.1: *Kinetics measurements for α-lytic protease mutants*

Protease	P_1	k_{cat}	K_M	k_{cat}/K_M	K_i
MMV	Ala	75	3.6	21000	67
MMV	Asn	0.69	20	34	
MMV	Cys	54	5.6	9600	
MMV	Gln	2.3	3.4	680	
MMV	Gly	12	35	330	
MMV	Leu	1.2	290	4.1	2000
MMV	Met	56	31	1800	
MMV	Ser	12	5.7	2000	
MMV	Thr	1.1	5.2	200	
MMV	Val	13	16	790	6.4
AMV	Ala	37	3.6	10000	64
AMV	Leu	87	0.77	110000	0.58
AMV	Met	120	0.33	350000	
AMV	Phe	130	0.40	310000	0.60
AMV	Val	3.4	1.1	3000	1.3
MAV	Ala	34	57	600	270
MAV	Leu			160	66
MAV	Met			980	
MAV	Phe	47	16	340	240

Protease	P_1	k_{cat}	K_M	k_{cat}/K_M	K_i
MAV	Val	10	29	340	210
MMA	Ala	74	4.9	15000	47
MMA	Leu	1.3	250	5.1	1400
MMA	Met	150	9.0	17000	
MMA	Val	5.6	20	280	30
SMV	Ala	26	4.4	6100	
SMV	Val				
SMV	Asn	39	6.3	6100	
SMV	Cys	190	1.9	99000	
SMV	Asp	0.29	75	3.8	
SMV	Gln	31	12	2600	
SMV	Glu	1.1	94	12	
SMV	His	5.6	3.3	1700	
SMV	Gly	1.2	20	60	
SMV	Ile	0.16	5.7	28	
SMV	Leu	66	3.7	18000	
SMV	Met	31	4.6	6700	
SMV	Phe	170	6.4	26000	
SMV	Ser	7.7	9.5	810	
SMV	Thr	24	2.2	11000	
SMV	Tyr	3.2	1.8	1700	
SMV	Val	93	4.2	22000	

Protease: one-letter codes for the substrate binding pocket amino acids at 192, 213, and 217A (*e.g.* MMV, the wild type enzyme, is Met192, Met213, and Val217A). P_1: refers to X in the indicator substrate succinyl-Ala-Ala-Pro-X-*p*-nitroanilide or in the boronic acid inhibitor methoxy-succinyl-alanyl-alanyl-prolyl-boro-X. k_{cat} (s^{-1}), K_M(mM), k_{cat}/K_M($s^{-1}M^{-1}$): kinetic constants determined from the best-fit lines of Lineweaver-Burke plots. K_i: inhibition constants (nM) were determined by competition with substrate. Substrate assays used 0.1→100mM substrate, 0.01→1.0 μM enzyme, in 100mM Tris-HCl, pH 8.0.

Table 7.2: *Comparison of experimental and calculated binding energies*

Protease	Substrate	Experimental	Simulation			P/T
			$w_{NB}=w_{solv}=1$	optimal w_{NB},w_{solv}	optimal weights+NGS	
MMV	Ala	0.00	0.00	0.00	0.00	P
MMV	Asn	3.82	-21.20	0.96	2.04	T
MMV	Cys	0.41	-1.35	-1.16	-0.89	P
MMV	Gln	2.05	41.46	3.03	2.55	P
MMV	Gly	2.46	-0.83	-0.02	2.38	P
MMV	Leu	5.05	83.10	2.30	3.67	P
MMV	Met	1.46	56.43	0.62	0.86	P
MMV	Ser	1.37	-26.43	0.89	1.90	T
MMV	Thr	1.46	-22.59	0.11	1.05	T
MMV	Val	1.91	-0.96	-0.37	0.11	T
AMV	Ala	0.41	-1.47	-0.48	-0.36	P
AMV	Leu	-0.95	11.15	-0.82	-0.82	P
AMV	Met	-1.64	6.55	-1.30	-0.93	T
AMV	Phe	-1.64	19.36	-1.49	-0.75	T
AMV	Val	1.09	-1.94	-0.82	-0.19	P
MAV	Ala	2.05	-1.77	-0.25	-0.17	P
MAV	Leu	2.87	42.30	1.03	2.50	T
MAV	Met	1.78	46.90	-0.15	-0.04	P
MAV	Phe	2.46	52.08	0.01	0.61	P

Protease	Substrate	Experimental	Simulation			P/T
			$w_{NB}=w_{solv}=1$	optimal w_{NB}, w_{solv}	optimal weights+NGS	
MAV	Val	2.46	3.25	-0.38	0.31	T
MMA	Ala	0.21	0.41	-0.23	-0.28	T
MMA	Leu	4.85	72.72	1.72	3.25	T
MMA	Met	0.10	45.40	0.02	0.42	T
MMA	Val	2.64	-0.67	-0.47	0.27	T
SMV	Ala	0.71	-2.99	-0.48	-0.36	P
SMV	Asn	0.68	-23.76	0.42	1.74	T
SMV	Asp	5.05	-15.82	2.73	3.84	T
SMV	Cys	-0.95	-4.58	-1.56	-0.98	T
SMV	Gln	1.23	-40.02	0.74	1.35	T
SMV	Glu	4.37	-37.93	2.69	3.65	T
SMV	Gly	3.41	-1.68	-0.18	2.21	P
SMV	His	1.50	5.35	1.22	1.76	P
SMV	Ile	3.82	132.05	2.99	3.39	P
SMV	Leu	0.00	9.42	-0.83	-0.85	P
SMV	Met	0.68	6.09	-1.26	-0.89	T
SMV	Phe	-0.13	18.11	-1.48	-0.74	P
SMV	Ser	1.91	-26.49	0.51	1.78	P
SMV	Thr	0.41	-24.48	-0.34	0.74	T
SMV	Tyr	1.50	20.98	-0.30	0.29	P
SMV	Val	0.00	-3.50	-0.82	-0.19	T

Protease: one-letter codes for the substrate binding pocket amino acids at 192, 213, and 217A (*e.g.* MMV, the wild type enzyme, is Met192, Met213, and Val217A). *Substrate*: refers to X in the indicator substrate succinyl-Ala-Ala-Pro-X-*p*-nitroanilide. *Experimental*: binding energies (Kcal/mole) calculated as described in methods. *w_{NB}=w_{solv}=1:* Calculated values using non-optimized parameters (w_{NB}=w_{solv}=1). *optimal w_{NB},w_{solv}:* Calculated values using optimized weights (w_{NB}=0.031, w_{solv}=1.98). *optimal weights + NGS:* Calculated values using fully optimized parameters (w_{NB}=0.031, w_{solv}=1.98, *scale for non-ground state energy term* = 5.7, *scale for* $\Delta\Delta G$ = 25.). *P/T:* Indicates whether the data point was included in the parameterization set (P) or in the testing set (T).

Table 7.3: *Importance of components to the model*

Modification to the model	r	$P(z\text{-}z_0)$
Standard model	0.85	—
Using non-bonded energies only	-0.09	$<10^{-8}$
Using solvation energies only	0.36	$<10^{-3}$
Using only the ground state rotamer combination	0.75	0.11
Using only the most common library rotamer at each site	-0.26	$<10^{-10}$
Using only ΔG(substrate), ΔG(complex)	0.66	0.02
Using only ΔG(enzyme), ΔG(complex)	0.09	$<10^{-6}$
Using only ΔG(substrate), ΔG(enzyme)	0.20	$<10^{-5}$
Using only ΔG(complex)	0.23	$<10^{-4}$

Binding energies were recalculated after slightly modifying the algorithm as described above. *r:* linear regression correlation coefficient, calculated between 40 calculated and experimental binding energies. $P(z\text{-}z_0)$: significance level of the difference between the Fisher z-transformations of r and r_0 (the standard model correlation) (Press *et al.*, 1986). The significance level indicates the likelihood that the modified model is actually better at predicting experimental binding energies than the standard model.

Table 7.4 *Activity of a designed protease with leucine and isoleucine substrates*

Enzyme	Substrate	k_{cat}/K_M $s^{-1}M^{-1}$	$\Delta G(exp)$ Kcal / mole	$\Delta G(calc)$ Kcal / mole
Wild type				
	Ala	21,000±700	0	0
	Leu	4.2± 0.05	5.05	3.67
	Ile	2.0± 0.04	5.47 (+0.42)	3.83 (+0.16)
Met→Val192				
	Leu	5320±210	0.82	0.85
	Ile	24±0.7	4.01 (+3.20)	7.90(+7.05)

Comparison of kinetic data and calculated vs. experimental binding energies for wild type and Met→Val192 mutant with either leucine or isoleucine substrates (suc-AAP-X-pNA, X=Leu or Ile). Experimental data for the Ala and Leu substrates + wild type enzyme have been reported previously (Bone et al., 1989b). Experimental and theoretical binding energies were calculated relative to the wild type + Ala substrate using equation 5. The difference between leucine and isoleucine binding for each enzyme is shown in brackets.

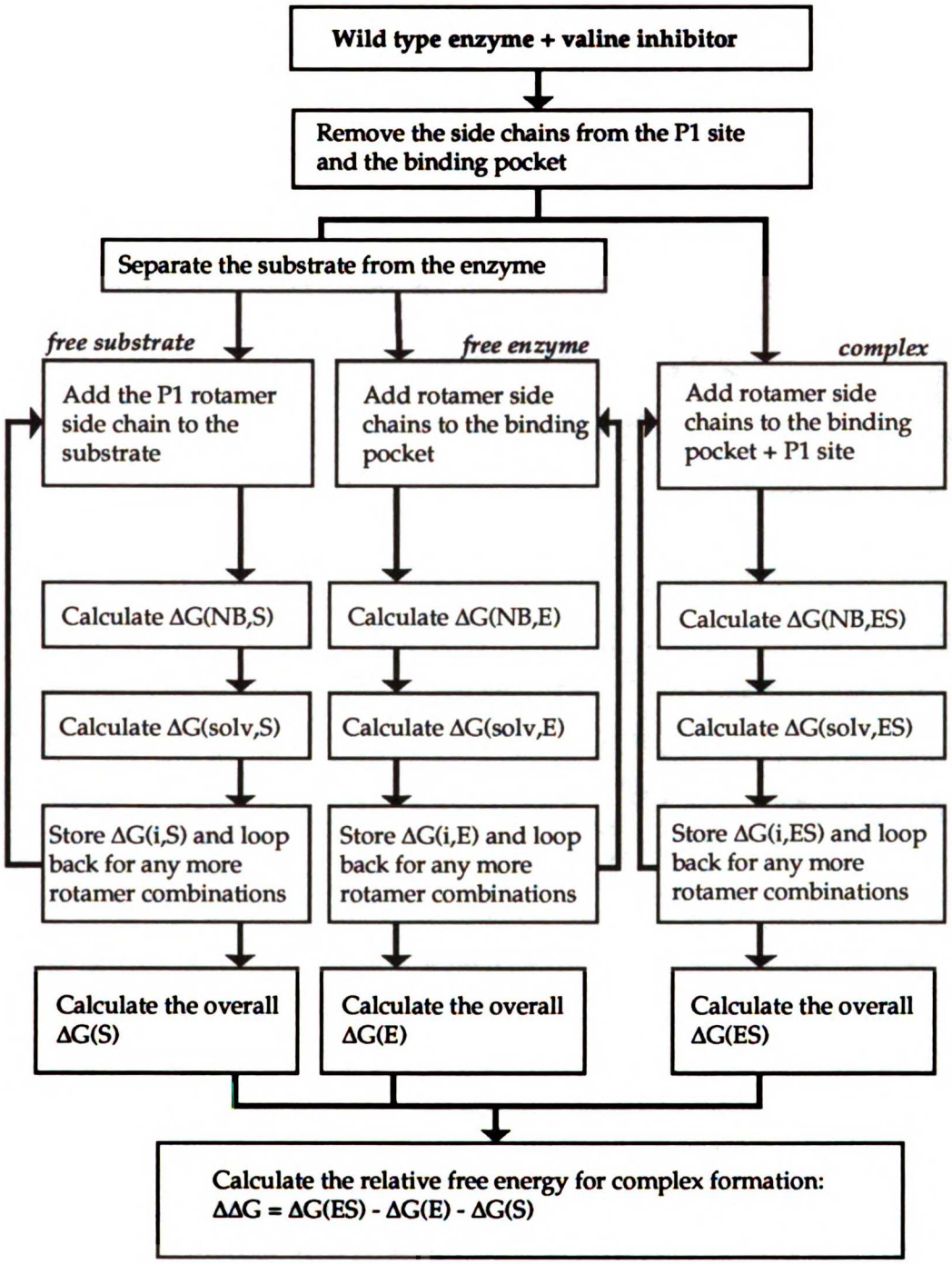


Figure 7.1. Algorithm for calculating relative binding energies. The flow chart shows the steps in calculating the relative binding energy for an enzyme-substrate complex. Details are given in METHODS.

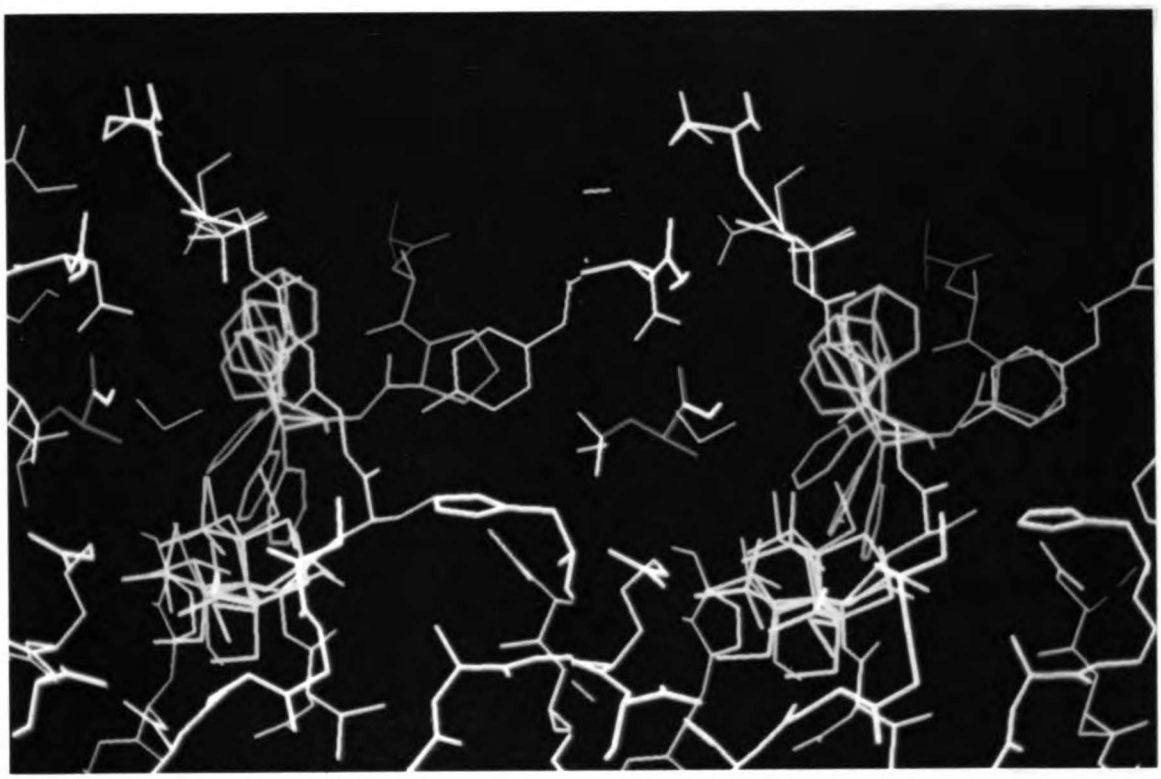


Figure 7.2. Sample rotamer model of α-lytic protease active site. All rotamers corresponding to the Met192→Ala protease mutant with the phenylalanine substrate bound are shown simultaneously. The side chain atoms for the enzyme residues 192 (red), 213 (yellow), and 217A (magenta), and the substrate P_1 residue (green) are shown. Conformation space is defined as the set of all rotamer combinations.

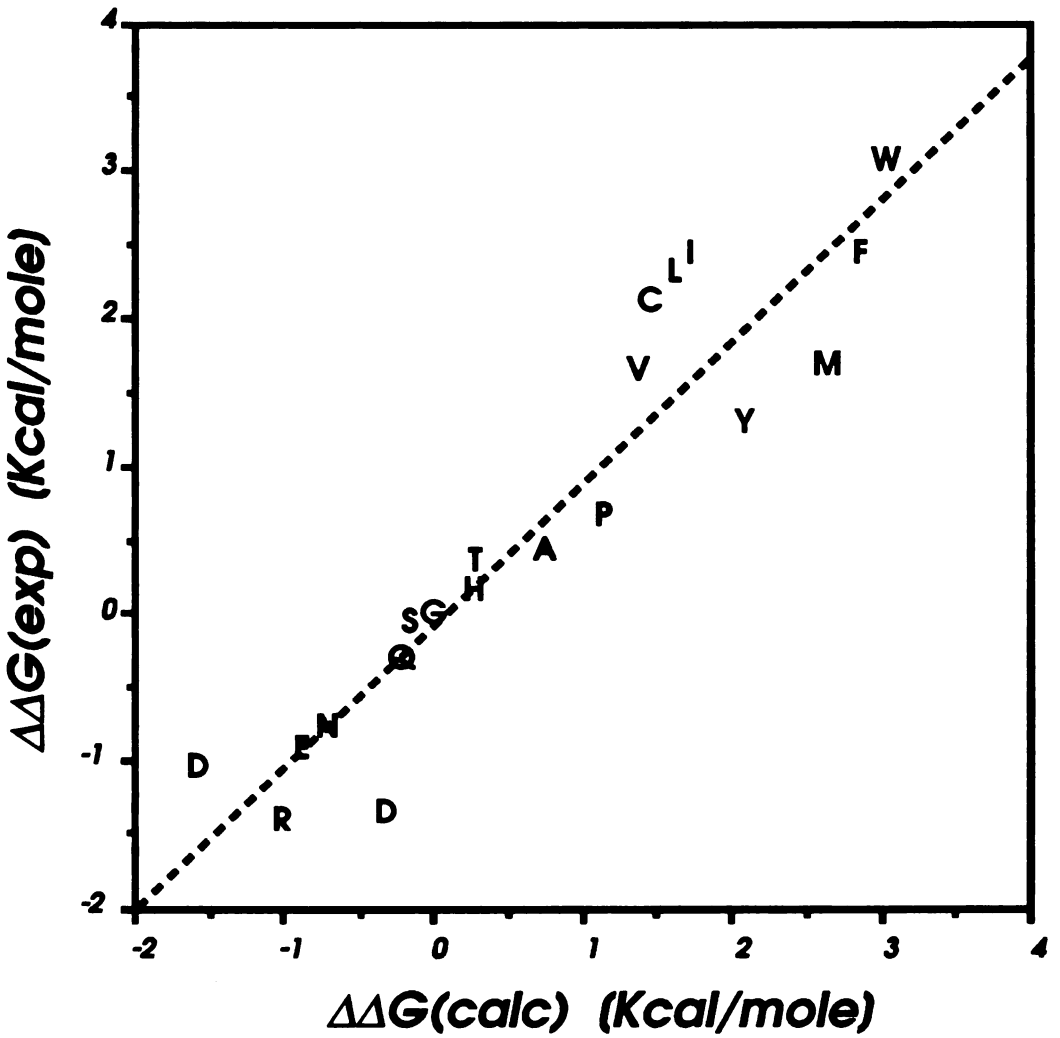


Figure 7.3. Parameterization of the solvation model. Five atomic solvation parameters (corresponding to carbon, non-charged nitrogen, non-charged oxygen, charged nitrogen/oxygen, and sulfur) were adjusted to optimize the fit to experimental *n*-octanol→water transfer free energies for 20 N-acetyl amino acid amides (Fauchere & Pliska, 1983). Residue 2 of α-lytic protease (in an extended conformation) was used to model the backbone of the blocked amino acid. Side chain atoms corresponding to all possible rotamers were added using the Ponder-Richards PROPAK program (1987b). The free energy for each rotamer was calculated as described in METHODS with either 1) all ASPs set to zero (corresponding to an *n*-octanol environment), 2) ASPs set to the optimal values (corresponding to an aqueous environment). The free energy of transfer for each amino acid was calculated as the difference between the Boltzmann-weighted sums over all rotamers for each of the two states. The experimental values relative to the glycine free energy are shown as a function of the calculated free energies relative to glycine.

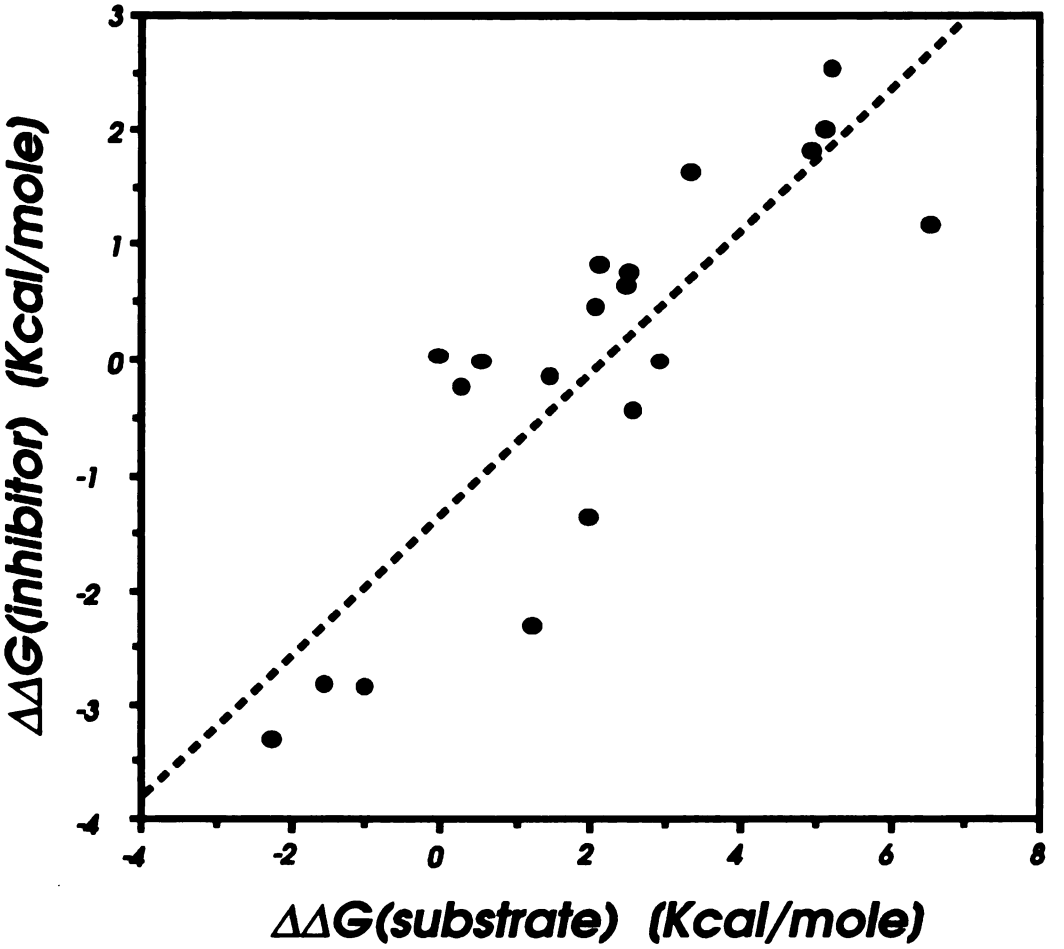


Figure 7.4. Boronic acid peptides as a model of the transition state. The binding energy of the boronic acid inhibitors Ala-Ala-Pro-X-boronic acid (calculated as $RTln(K_i)$) using the K_is reported in table 7.1) are plotted as a function of the corresponding binding energies for the substrates Ala-Ala-Pro-X-*p*-nitroanilide (calculated as $-RTln(k_{cat}/K_M)$). Both values are normalized by the subtracting the binding energy calculated for the alanine inhibitor/substrate.

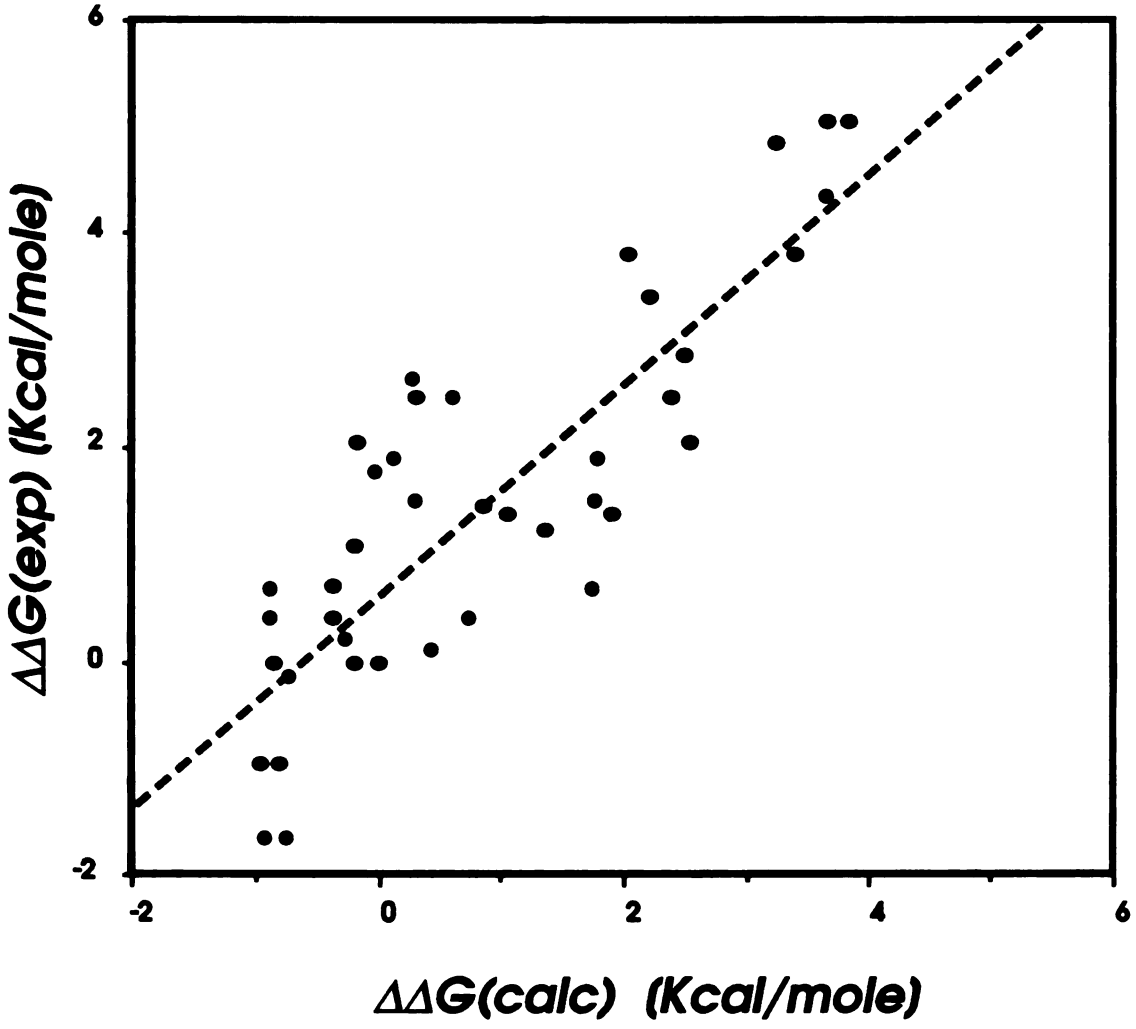


Figure 7.5. Calculated vs. experimental binding energies. Binding energies calculated for each of the enzyme-substrate combinations listed in table 7.2 (both parameterization and testing data sets) are shown as a function of the experimentally-observed binding energy (using the wild-type enzyme with the alanine substrate as the reference state). The solid line is the least-squares best fit line (experiment = 0.99 × calculated + 0.68).

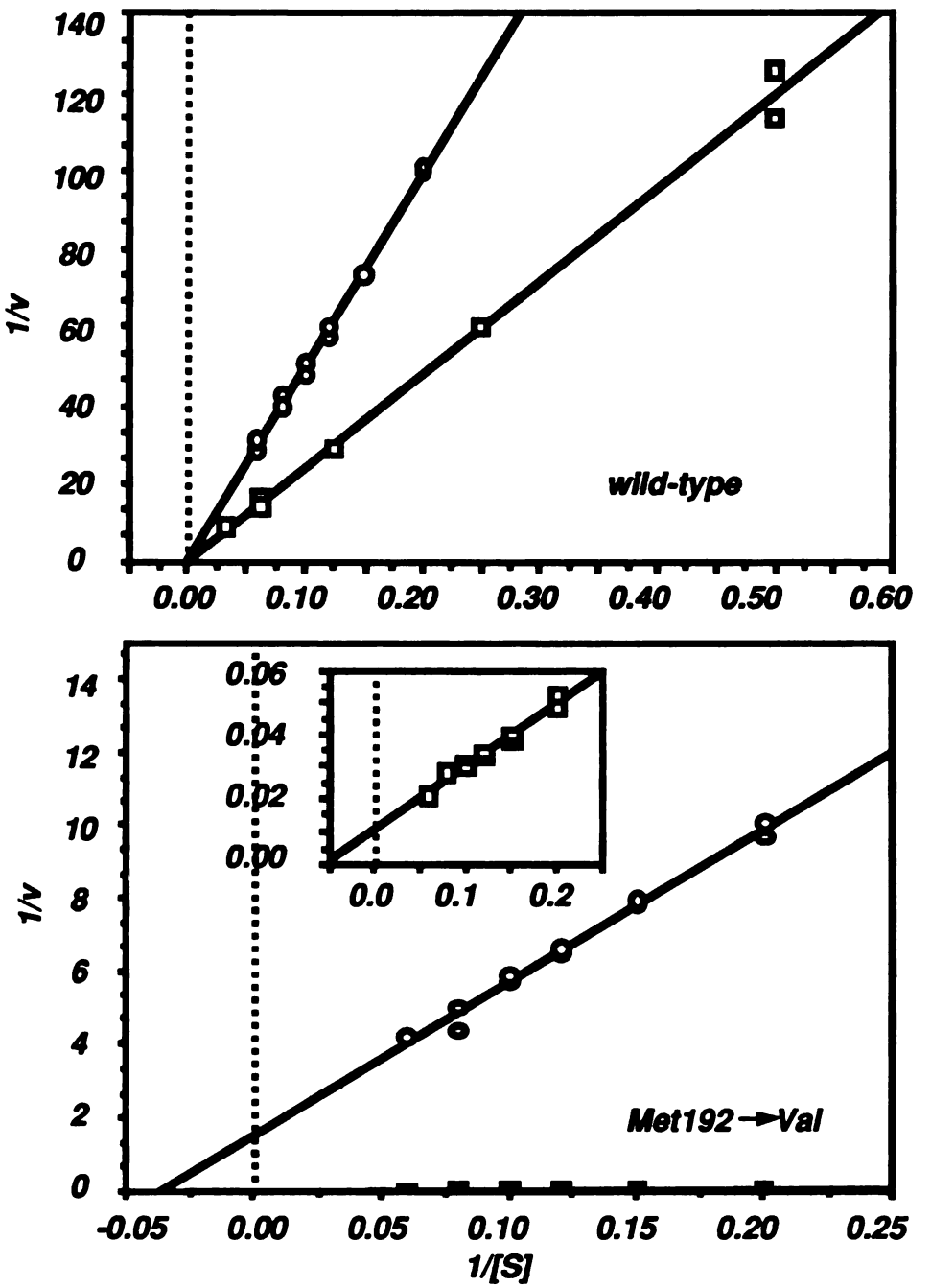


Figure 7.6. Kinetics measurements for an engineered protease. Kinetic constants were determined for both the wild type and Met192→Val mutant for succinyl-alanyl-alanyl-prolyl-X-*p*-nitroanilide substrates with X = leucine (squares) or isoleucine (circles). For wild type enzyme only k_{cat}/K_M could be accurately measured: with the leucine substrate, k_{cat}/K_M = 4.2±0.05 $M^{-1}s^{-1}$; with the isoleucine substrate, k_{cat}/K_M = 2.0±0.04 $M^{-1}s^{-1}$. For the Met→Val192 mutant with the leucine substrate, K_M = 16±1.5 mM, k_{cat} = 86±7 s^{-1}, k_{cat}/K_M = 5320±210 $M^{-1}s^{-1}$; for the isoleucine substrate, K_M = 26±3 mM, k_{cat} = 0.64±0.07 s^{-1}, k_{cat}/K_M = 24±0.7 $M^{-1}s^{-1}$. Results are summarized in table 7.4.

References

Bash, P.A., Singh, U.C., Langridge, R. & Kollman, P.A. (1987). Free energy calculations by computer simulation. *Science*, **236**, 564-568.

Bone, R., Shenvi, A.B., Kettner, C.,A. & Agard, D.A. (1987). Serine protease mechanism: structure of an inhibitory complex of alpha-lytic protease and a tightly bound peptide boronic acid. *Biochemistry*, **27**, 7609-7614.

Bone, R., Silen, J.L., Agard, D.A. (1989a). Structural plasticity broadens the specificity of an engineered protease. *Nature*, **339**, 191-195.

Bone, R., Frank, D., Kettner, C.A., Agard, D.A. (1989b). Structural analysis of specificity: alpha-lytic protease complexes with analogues of reaction intermediates. *Biochemistry*, **28**, 7600-7609.

Bone, R., Sampson, N.S., Bartlett, P.A. & Agard, D.A. (1990). Crystal structures of alpha-lytic protease complexes with irreversibly bound phosphonate esters. *Biochemistry*, in press.

Caldwell, J.W., Agard, D.A., Kollman, P.A. (1990). Free energy calculations on binding and catalysis by alpha-lytic protease: the role of substrate size in the P_1 pocket. *Proteins*, in press.

Dorovska-Taran, V, Momtcheva, R., Gulubova, N. & Martinek,K. (1982). The specificity in the elementary steps of alpha-chymotrypsin catalysis: A temperature study with a series of N-acetyl-L-amino acid methyl esters. *Biochim. Biophys. Acta*, **702**, 37-53.

Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199-203.

Fauchere, J.L. & Pliska, V. (1983). Hydrophobic parameters π of amino-acid side chains from the partitioning of N-acetyl-amino-acid-amides. *Eur. J. med. Chem. - Chim. ther.* **18**, 369-375.

Fersht, A.R. (1985a). *Enzyme structure and mechanism*, 2nd edition. W.H. Freeman & Co. New York, NY.

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. (1985b). Hydrogen-bonding and biological specificity analysed by protein engineering. *Nature*, **314**, 235-238.

James, M.N.G., DelBaere, L.T.J. & Brayer, G.D. (1978). Amino acid sequence alignment of bacterial and mammalian pancreatic serine proteases based on topological equivalences. *Can. J. Biochem.* **56**, 396-402.

Kettner, C.A., Bone, R., Agard, D.A. & Bachovchin, W.W. (1988). Kinetic properties of the binding of alpha-lytic protease to peptide boronic acids. *Biochemistry*, **27**, 7682-7688.

Lee, B. & Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.

Matthews, D.A., Alden, R.A., Birktoft, J.J. & Kraut, J. (1975). X-ray crystallographic study of boronic acid adducts with subtilisin BPN' (Novo). *J. Biol. Chem.* **250**, 7120-7126.

McCammon, J.A., Harvey, S.C. (1987). *Dynamics of proteins and nucleic acids*, p. 29, Cambridge University Press, New York, NY.

Naray-Szabo, G. (1989). Quantitative estimation of activities of mutant enzymes. *Catalysis Letters*, **2**, 185-190.

Novotny, J., Bruccoleri, R.E., & Saul, F.A. (1989). On the attribution of binding energy in antigen-antibody complexes McPC603, D1.3, HyHEL-5. *Biochemistry*, **28**, 4735-47749.

Ponder,J.W. & Richards, F.M. (1987a). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791.

Ponder, J. (1987b). PROPAK program.

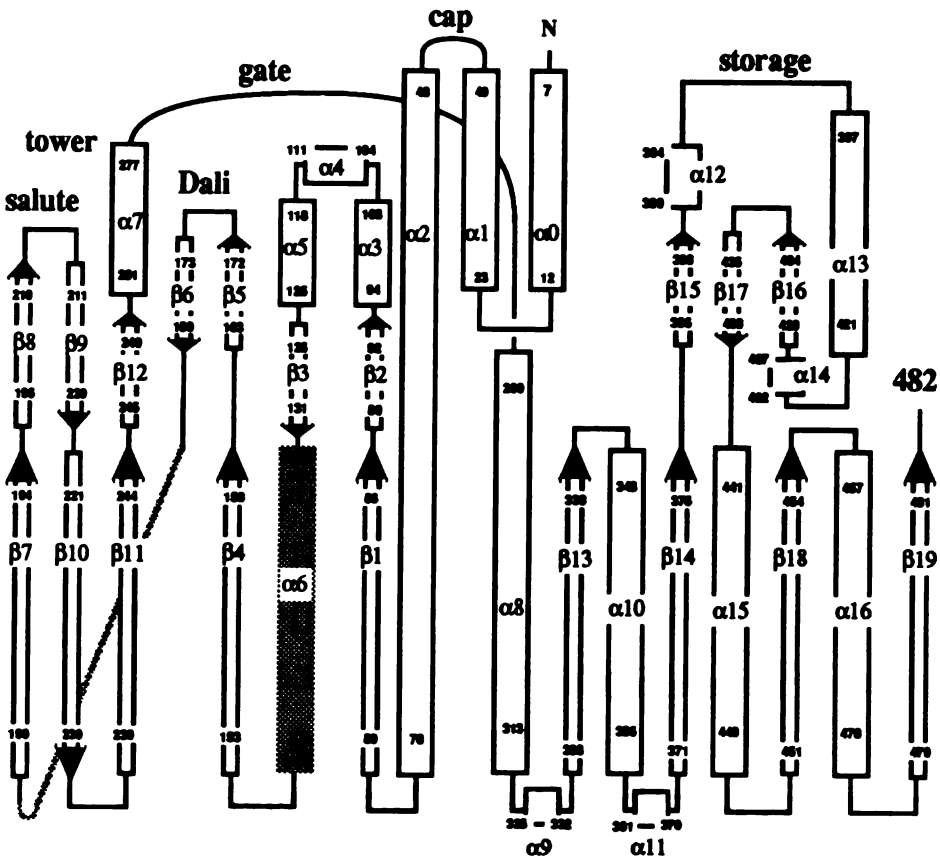
Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1986). *Numerical recipes: the art of scientific computing*, p. 486, Cambridge University Press, New York, NY.

Rao, S.N., Singh, U.C., Bash, P.A. & Kollman, P.A. (1987). Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature*, **328**, 551-554.

Shrake, A. & Rupley, J.A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and Insulin. *J. Mol. Biol.* **79**, 351-371.

Warshel, A., Sussman, F., Hwang, J.K. (1988). Evaluation of catalytic free energies in genetically modified proteins. *J. Mol. Biol.* **201**, 139-159.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765-784.


cap
N
gate
storage
tower
salute
Dali
α12
α13
α14
482
α7
α4
α5
α3
α2
α1
α0
β6
β5
β15
β17
β16
β8
β9
β12
β3
β2
β7
β10
β11
β4
β1
α6
α8
β13
α10
β14
α15
β18
α16
β19
α9
α11

Chapter 8 :

Development of a genetic screen for protease activity based upon pro-inducer conversion

Abstract

Genetic analysis of mutant proteins has played a key role in many recent structure—function studies. The application of this technique to the study of enzymes requires that altered enzymatic activity be easily detectable *in situ*. We have developed a screen for mutant proteases in which proteolytic activity triggers the expression of a reporter enzyme that can be easily detected using standard chromogenic substrates. Our method is based on the *dsdA* [D-serine deaminase] operon of *E. coli*. Although normally inactive, this operon can be strongly induced by low concentrations of D-serine or D-serine-containing dipeptides. A longer peptide containing D-serine at the C-terminus will not act as an inducer unless a protease is present to cleave it, thereby releasing D-serine or X-D-serine. Active protease can be detected if the *lac Z* gene (β-galactosidase) is placed under the control of the *dsdA* promoter. Coupling enzymatic activity to the induction of a reporter gene via pro-inducer conversion may be a generally useful strategy for studies with enzymes that cannot be readily detected *in situ*.

Introduction

Genetic screens and selections are extremely powerful tools for probing the structural basis of protein function. Site-directed mutagenesis of the lac repressor (Lehming, 1987) and the trp repressor (Bass, 1988) has led to a much better understanding of how these proteins bind selectively to their target operator sequences. Both of these studies were possible because altered repressor activities could be easily detected within a large library of mutants. Recently there has been much interest in using the serine proteases as model systems to explore the determinants of enzyme specificity (Estell, 1986; Wells, 1987; Evnin, 1990). We have focused on α-lytic protease, a small bacterial serine protease homologous to the trypsin family of mammalian enzymes. By individual mutation of residues lining the binding pocket, it has been possible to obtain highly active enzymes with significantly altered patterns of substrate specificity (Bone,1989). An important goal is to examine the ability of combinatorial binding pocket mutations to create unique enzyme specificities. To enable these studies, we have sought to create a protease screen / selection that would allow the same in depth structure-function analysis that has been possible with the lac and trp repressors (Lehming, 1987; Bass, 1988).

We now describe a sensitive method for screening large libraries of mutant enzymes on the basis of enzymatic activity towards a well-defined substrate. This screen has been implemented for our studies of mutant proteases, although in principle it should be possible to adapt it for studies with other types of hydrolytic enzymes.

Methods

Figure 8.1 shows the general scheme of the genetic screen. The screen functions by linking the catalytic activity of the enzyme of interest to the induction of a reporter

enzyme that can be easily detected. To accomplish this, a pro-inducer for the reporter enzyme is synthesized. This pro-inducer is unable to induce reporter enzyme expression but can be converted by the enzyme of interest into an active form.

To develop a screen for proteases with altered proteolytic activities, we have utilized the D-Serine deaminase operon (*dsdA*). McFall and coworkers have thoroughly characterized the properties of this system in *E. coli* (see review, McFall, 1987). D-Serine deaminase is very poorly expressed under normal growth conditions (≈35 molecules / cell). The addition of low levels of D-serine (<1μM), however, significantly enhances expression of the deaminase. With higher concentrations of D-serine (>10μM), expression can be increased 3000- to 5000-fold and the deaminase then makes up ≈1% of the total cell protein. While D-serine and D-serine-containing dipeptides lead to induction of the *dsdA* gene, we have shown that longer D-serine-containing peptides will not. To develop a protease screen, we designed a peptide pro-inducer with the sequence L-Ala-L-Ala-L-Pro-L-Ala-L-Ala-D-Ser. We showed by reverse-phase HPLC (Figure 8.2) that wild-type α-lytic protease would selectively cleave this peptide to yield the tetrapeptide Ala-Ala-Pro-Ala and the dipeptide L-Ala-D-Ser (tests with the pentapeptide Ala-Ala-Pro-Ala-D-Ser, showed that α-lytic protease was unable to release D-ser). Using *E. coli* cells that have the *lac Z* gene inserted into the coding region for D-serine deaminase and carry an expression vector for α-lytic protease, we made expression of β-galactosidase (which can be easily detected with chromogenic substrates) dependent upon the presence of an active protease.

There are many advantages to such a screen. The pro-inducer can be easily made using a solid-phase peptide synthesizer, making it feasible to generate a family of pro-inducers to assay for different types of protease activity (using the series of Ala-Ala-Pro-X-Ala-D-Ser peptides). D-Serine induces *dsdA* expression at submicromolar concentrations and therefore only small amounts of pro-inducer need be provided (Heincz, 1984). At substrate concentrations significantly below the K_M, the level of reporter gene

induction should be directly related to k_{cat}/K_M (and thus a function of $\Delta G^{\dagger}$, the reaction transition state stabilization energy). Hexapeptides are chemically stable and cannot be transported into cells — thus the pro-inducer alone should give a very low background of induction. D-Serine is actively concentrated inside bacterial cells with a high-affinity K_M for transport (Cosloy, 1973). Therefore, once the pro-inducer is cleaved, the inducer should be taken up locally and act only on the colony producing the active protease. The traits of the *dsdA* operon listed above (high sensitivity, stability of the pro-inducer, active transport of inducer) are common to several other *E. coli* operons and could be used in other adaptations of the screen.

To implement the screen, we required an inducible expression vector for α-lytic protease and an engineered cell line in which the expression of β-galactosidase is controlled by the *dsdA* operon. Dr. E.M. McFall kindly provided us with cell line EM111-1 (*dsdA*::Mu d(*lac* Ap^r) *dsdC*$^+$). This cell line has a promoterless *lac Z*-containing Mu transposon inserted into the coding region of the *dsdA* gene. Addition of D-serine to these cells induces β-galactosidase expression, as shown by growth on plates containing 5-bromo-4-chloro-3-indolyl-beta-D-galactoside (X-gal). 0.1—1000 μM D-serine causes the colonies to turn blue, with a graded response covering the entire concentration range. The *E coli. lac Z* gene is non-functional, thus β-galactosidase expression is completely independent of the lac promoter (*i.e.* is not induced by IPTG).

We designed an expression vector for α-lytic protease based on the pHSe5 vector (Muchmore, 1989). This vector has been used for the high level expression of T4 lysozyme mutants and was chosen because protein expression (driven by tandem *tac* promoters) can be easily controlled by the addition of IPTG. The pHSe3 vector can be selected with ampicillin but since the EM111-1 cells are already ampicillin-resistant, we replaced the plasmid Ap^r gene with a fragment conferring Kanamycin-resistance (using the

Pharmacia Kan^r genblock, inserted into the Pst I sites on pHSe3). The coding region of lysozyme was removed from this plasmid using the Nde I and Xba I restriction sites, and replaced by a Mae I — Xba I fragment containing the gene for α-lytic protease (derived from pDA30). EM111-1 cells were transformed with pDA50 and protease activity assays of cell culture supernatants showed that these cells secreted active α-lytic protease. Protease activity was shown to be strongly controlled by IPTG (Figure 8.3).

There are convenient restriction sites have been engineered within the protease coding region to allow the generation of random or site-specific mutants (Silen, 1989). For this study, we have used the Sty I and Xba I sites to swap fragments containing the coding region for the substrate binding pocket between different plasmid constructs.

Results

To determine whether active protease could be detected using this scheme, EM111-1 cells transformed with the protease expression vector were grown in selective media (L-broth + Kanamycin + Carbenicillin) containing 500 μM pro-inducer Ala-Ala-Pro-Ala-Ala-D-Ser, and 0.0—1.0 mM IPTG (Figure 8.4). Because α-lytic protease is poorly expressed at higher temperatures (Silen, 1989), cells were incubated at 23° for several days. Cells grown in media lacking IPTG (therefore not expressing α-lytic protease) remained white following the addition of X-gal. After several days, the addition of 1.0 mM IPTG (inducing high protease levels) caused the cell culture to turn an intense dark blue color following X-gal addition. Control EM111-1 cells transformed with the pDA48 plasmid (expressing T4 lysozyme rather than α-lytic protease) showed no signal with or without the addition of IPTG.

α-lytic protease was shown by HPLC to release the dipeptide Ala-D-Ser upon cleavage of the hexapeptide Ala-Ala-Pro-Ala-Ala-D-Ser. Previous studies suggested that dipeptides do not directly activate the *dsdA* operon, but that activation occurs as a result of the release of D-serine following intracellular hydrolysis of the dipeptide (E.M. McFall, personal communication). We questioned whether amino acids other than alanine at the P1' position (Ala-Ala-Pro-Ala-X-D-Ser) would improve the screen by either reducing non-α-lytic protease-catalyzed peptide cleavage, or by increasing the rate of intracellular dipeptide hydrolysis. A series of peptides, X=Ala, Gly, Leu, Met, Phe, Asp, or Lys, were incubated overnight in the presence or absence of purified α-lytic protease and subsequently added to log-phase EM111-1 cells (not carrying the α-lytic protease expression vector). As shown in Figure 8.5, α-lytic protease-independent induction of β-galactosidase varies considerably for the different peptides. Alanine and glycine give the largest non-specific signal, while larger amino acids and charged amino acids give almost no detectable α-lytic protease-independent induction. The strength of the α-lytic protease-dependent signal also varied considerably between peptides, being optimal for the Met peptide (Ala-Ala-Pro-Ala-Met-D-Ser). Differences in the size of the signal could be due to different rates of cleavage by α-lytic protease, or due to differences in the *dsdA* induction potential of the resulting dipeptide. Previous studies have suggested that α-lytic protease shows relatively little substrate specificity at the P1' position (Delbaere, 1981), arguing for the latter explanation. It is interesting to note that aminopeptidases synthesized by E. coli are somewhat Met-specific (since Met must be cleaved from many newly-synthesized proteins) (Ben-Bassat, 1987); it is therefore not surprising that the Met-D-Ser peptide can efficiently induce the *dsdA* operon.

Since protease activity could be easily detected *in situ*, we tested whether the screen could distinguish between good and bad substrates for an enzyme. The preferred substrates for wild type α-lytic protease have been determined using purified enzyme and

p-nitroanilide tetrapeptide substrates. α-Lytic protease shows strong selectivity for substrates at the P1 position (the amino acid on the N-terminal side of the scissile peptide bond, the fourth residue in our hexapeptide pro-inducers) (Delbaere, 1981). For instance, k_{cat}/K_M for the alanine substrate (Ala-Ala-Pro-Ala-*p*-nitroanilide) is 21,000 Ms^{-1}, whereas that for the phenylalanine substrate (Ala-Ala-Pro-Phe-*p*-nitroanilide) is only 0.38 Ms^{-1} (Bone, 1989). We synthesized a pro-inducer in which the alanine at the fourth position was replaced by a phenylalanine. EM111-1 / pDA50 cells were then grown on media containing either the alanine or the phenylalanine pro-inducer, together with IPTG and X-gal. As shown in Figure 8.6, cells respond well to the alanine pro-inducer but show virtually no response to the phenylalanine pro-inducer.

To function as a useful screen, one must be able to select individual cells from the entire population on the basis of activity. While the current tests in liquid culture demonstrate that active protease can be detected, it would be impossible to identify individual mutants since active protease from one cell is secreted into solution where it is actively mixed with all other cells. To avoid this problem, cells must be grown on solid media as individual colonies. Tests of the screen on solid media have been quite variable; occasionally a strong difference between protease-expressing and non-expressing cells can detected whereas the same experiment at other times shows no significant signal.

One possible explanation for this result is given by the observation that stationary phase cells to not induce the *dsdA* operon in response to D-serine. A 1 ml culture of EM111-1 cells was grown for two days under selective conditions in L-broth (23°, shaking air incubator). 5 μl of the cell culture was removed after two days and diluted by L-broth to start a new culture. D-serine (500 μM) was simultaneously added to both the old and the fresh culture. After one day, X-gal was added to the cells and the color change was recorded. D-serine strongly induced β-galactosidase expression by the fresh culture, while the old culture showed non-specific induction and no D-serine-dependent induction.

Except for cells lying at the periphery, a bacterial colony contains cells which are no longer actively growing and might therefore be expected to behave like a stationary-phase culture with regard to *dsdA* induction. If this is the case, by searching for conditions which allow *dsdA* induction in liquid culture stationary phase cells, one may be able to improve the functioning of the screen with cells on solid media.

Discussion

The above experiments demonstrate that the screen can successfully identify cells expressing active protease. Following optimatization of the media and growth conditions on solid agar plates, the screen should be directly applicable to structure—function studies for α-lytic protease, especially for experiments designed to probe the determinants of substrate specificity.

The sensitivity of the assay may be easily adjusted by controlling levels of the reporter substrate or of the pro-inducer. With large amounts of pro-inducer present, barely active mutants should be distinguished from completely inactive mutants. Conversely, small amounts of pro-inducer should allow a signal for only the most highly active enzymes. In addition to D-serine control, the *dsdA* operon is also regulated by the cyclic AMP-cAMP binding protein complex (McFall, 1973). In *crp*$^-$ (cAMP receptor protein) or *cya*$^-$ (adenylate cyclase) cells, the D-serine-based induction of deaminase is somewhat attenuated. It is possible that this effect can also be used to alter the sensitivity of the screen.

The screen can be readily converted into a selection by using an antibiotic resistance-encoding gene rather than a reporter gene. If Cm^r is used instead of the *lac Z* gene, proteolytic activity should give rise to chloramphenicol resistance and cells should

then survive selection with this antibiotic. Alternatively, it is possible to use the expression of β-galactosidase as the basis for a selection by growing EM111-1 / pDA50 cells on minimal lactose media. Cells which secrete active protease induce *lac Z* expression and are thus able to utilize the lactose and grow. Cells without active protease are carbon-starved and tend to grow much more slowly. Preliminary tests of this selection are encouraging although the background growth of non-induced cells remains a technical problem (after several days, a small number of cells seem to lose their lacZ$^-$ phenotype and start growing at a normal, unrestricted rate).

It should be possible to adapt this scheme for other types of enzymatic activity using different pro-inducers. For instance, by coupling IPTG to an oligosaccharide, active glycosidases could be selected on the basis of *lac Z* induction. The range of enzymatic activities which can be screened is currently limited only by the variety of inducible promoters available in *E. coli*.

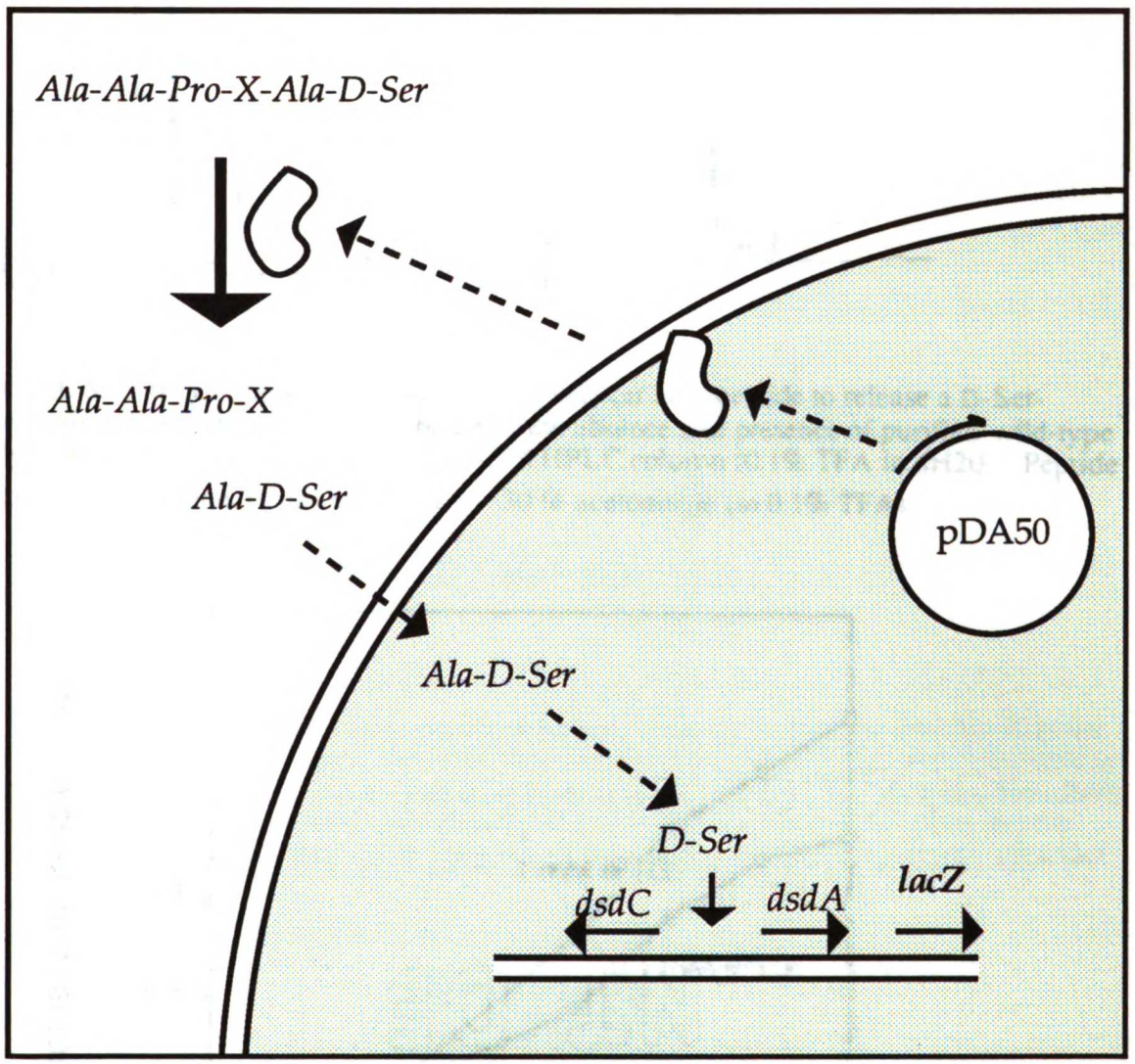


FIGURE 8.1. Scheme for the protease screen. A detailed description appears in the text.

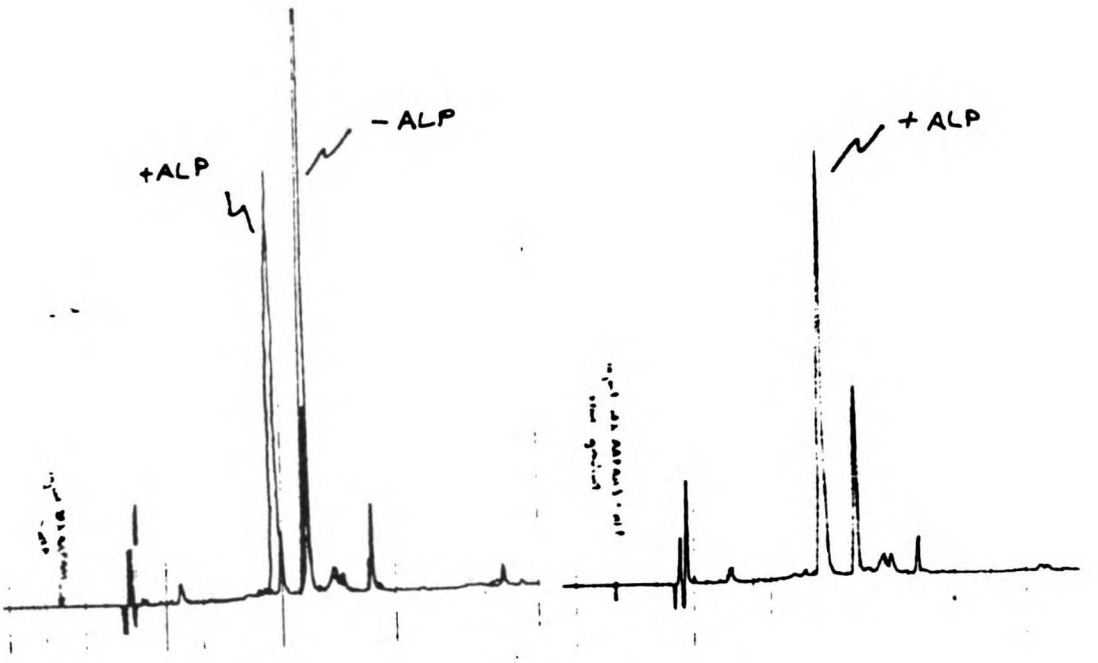


FIGURE 8.2. α-lytic protease cleaves the pro-inducer hexapeptide to release a D-Ser-containing dipeptide. Peptide incubated in the absence and presence of purified wild-type a-lytic protease was applied to a Vydac C18 HPLC column (0.1% TFA in dH20). Peptide was eluted by a linear gradient of 0 % → 30 % acetonitrile (in 0.1% TFA).

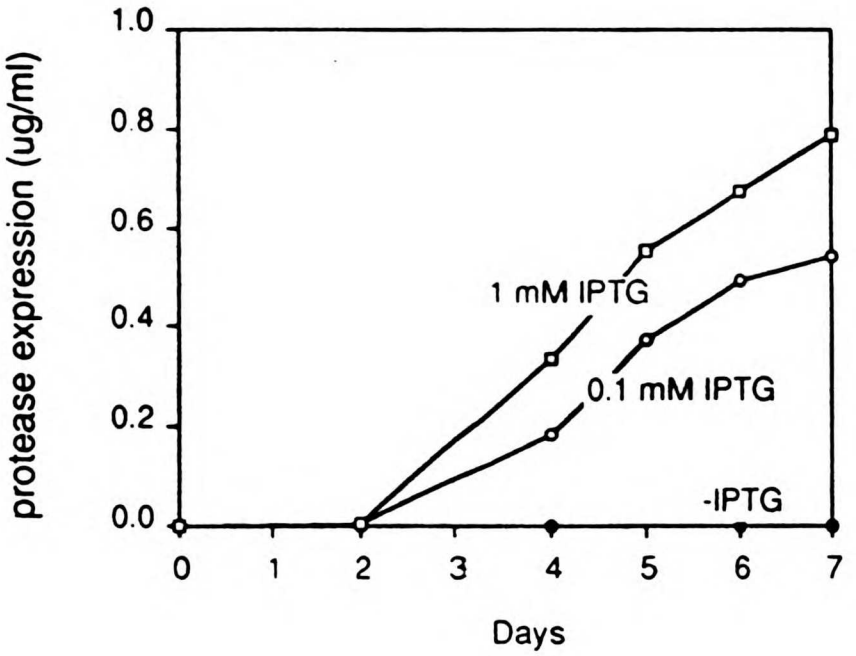


FIGURE 8.3. α-lytic protease expression. The expression of α-lytic protease was measured using the substrate Ala-Pro-Ala-p-nitroanilide for cell cultures grown in the presence or absence of IPTG.

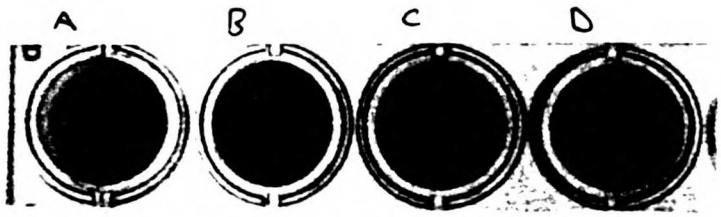


FIGURE 8.4 EM111-1 / pDA50 cells were grown on rich media containing carbenicillin, kanamycin, X-gal, with no (a) or with 500 µM ALA-pro-inducer (b,c), and with no (a,b) or with 1mM (c) IPTG. Control cells incubated with 500 µM D-serine (d) clearly induce β-galatosidase expression. Cells expressing α-lytic protease (+IPTG) show a clear signal as compared to non-induced (-IPTG) cells.

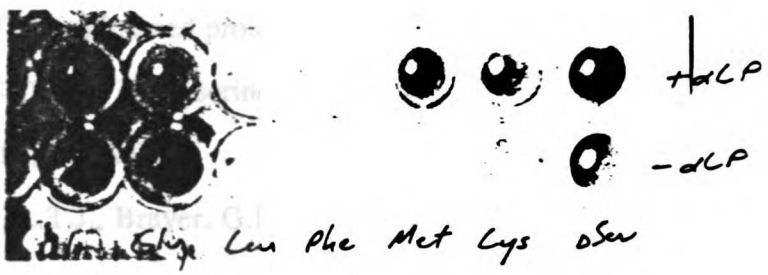


FIGURE 8.5. Specific- and non-specific induction of *dsdA* by Ala-Ala-Pro-Ala-X-D-Ser peptides (X = Ala, Gly, Leu, Met, Phe, Asp, or Lys). 10mM peptides were incubated at room temperature for 12 hours with or without α-lytic protease (10 nM) and subsequently added to log phase EM111-1 cells growing in a shaking incubator at 23° (final peptide concentration = 500 µM). After 1 day, X-gal was added and the color change recorded.

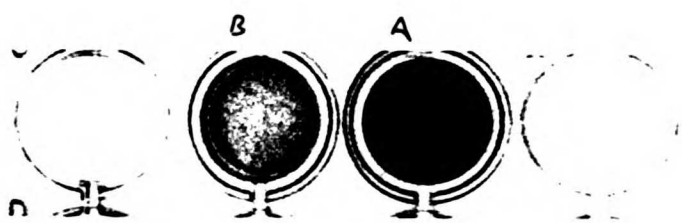


FIGURE 8.6. EM111-1 cells expressing α-lytic protease were grown in L-broth containing 1 mM IPTG, X-gal, and either 500 µM ALA-pro-inducer (a) or PHE-pro-inducer (b).

References

Bass, S., Sorrells, V., & Youderian (1988). Mutant Trp repressors with new DNA-binding specificities. *Science.* **242**, 240-245.

Ben-Basset, A., Bauer, K., Chang, S.Y., Myambo, K., Boosman, A., & Chang, S. (1987). Processing of the initiation methionine from proteins: properties of the Escherichia coli methionine aminopeptidase and its gene structure. *J. Bacteriol.* **169**, 751-757.

Bone, R., Silen, J.L., & Agard, D.A. (1989). Structural plasticity broadens the specificity of an engineered protease. *Nature*, **339**, 191-195.

Cosloy, S.D. (1973). D-serine transport system in Escherichia coli K-12. *J. Bacteriol.* **114**, 685-694.

Delbaere, L.T.J., Brayer, G.D., & James, M.N.G. (1981). *Eur. J. Biochem.* **120**, 289-294.

Estell, D.A., Graycar, T.P., Miller, J.V., Powers, D.B., Burnier, J.P., Ng, P.G., & Wells, J.A. (1986). Probing steric and hydrophobic effects on enzyme-substrate interactions by protein engineering. *Science.* **233**, 659-663.

Evnin, L.B., Vasquez, J.R., & Craik, C.S. (1990). Substrate specificity of trypsin investigated by using a genetic selection. *Proc. Natl. Acad. Sci, USA.* **87**, 6659-6663.

Heincz, M.C., Bornstein, S.M., & McFall, E. (1984). Purification and characterization of D-serine deaminase activator protein. *J. Bacteriol.* **160**, 42-49.

Lehming, N., Surtains, J., Niemoller, M., Genenger, G., Wilcken-Bergmann, S., & Muller-Hill, B. (1987). The interaction of the recognition helix of lac repressor with lac operator. *EMBO Journal.* **6**, 3145-3153.

McFall, E. (1973). Role of adenosine 3',5'-cyclic monophosphate and its specific binding protein in the regulation of D-serine deaminase synthesis. *J. Bacteriol.*, **113**, 781-785.

McFall, E. (1987). The D-serine deaminase operon. in Escherichia coli and Salmonella typhimurium: Cellular and molecular biology, Vol. 2, eds. Ingraham, J.L., Low, K.B., Magasanik, B., Schaechter, M., & Umbarger, H.E., American Society for Microbiology, Washington, D.C. 1520-1526.

Muchmore, D.C., McIntosh, L.P., Russell, C.B., Anderson, D.E., & Dahlquist, F.W. (1989). Expression and nitrogen-15 labelling of proteins for proton and nitrogen-15 nuclear magnetic resonance. *Meth. Enzymol.* **177**, 44-73.

Silen, J.L., Frank, D., Fujishige, A., Bone, R., & Agard, D.A. (1989). *J. Bacteriol.* **171**, 1320-1325.

Wells, J.A., Powers, D.B., Bott, R.R., Graycar, T.P., & Estell, D.A. (1987). *Proc. Natl. Acad. Sci., USA.* **84**, 1219-1223.

Chapter 9 :

Modelling side chain conformation for homologous proteins using an energy-based rotamer search

Abstract

We have developed a computational method for accurately predicting the conformation of side chain atoms when building a protein structure from a known homologous protein. A library of rotamers is used to model the side chains, allowing 5-6 different conformations per residue on average. Local sites of adjacent side chains are defined throughout the protein, and all combinations of side chain rotamers are evaluated within each site using a molecular mechanics force field enhanced by the inclusion of a solvation term. At each site, the lowest energy combination of side chains is identified and added onto the fixed protein backbone. A series of test cases using the refined x-ray structure of α-lytic protease has shown that 1) the force field can correctly predict up to 90% of side chain rotamers, 2) the assumption of side chain rotamer geometry is usually a very good approximation, and 3) the complete combinatorial conformation search is able to overcome local minima and identify the lowest energy rotamer set for the protein in the absence of a starting bias to the correct structure. Tests with several pairs of homologous proteins have shown that the algorithm is quite successful at predicting side chain conformation even when the protein backbone used to generate side chain positions is distorted. The average r.m.s. deviation of predicted side chain atoms rises from 0.73 Å (in a test case with the correct backbone) to only 1.95 Å (in a test case with <35% homology). The high accuracy of this method suggests that it may be a useful automated tool for modelling protein structure.

Introduction

Several different approaches have been developed to predict the structure of a protein using its amino acid sequence. The most successful class of prediction techniques uses the known structure of an homologous protein as a starting point for modeling the unknown structure (Blundell *et al.*, 1987). As the number of protein crystal structures grows, it becomes increasingly likely that an unsolved protein will have some homology to a previously-solved structure. In the future, therefore, it may be possible to predict the tertiary structure of many proteins using the database of solved structures. The development of an accurate, homology-based modelling algorithm is thus likely to have general use in solving the protein folding problem.

There are three major unsolved aspects to the problem of homology-based modelling: 1) constructing the alignment of a sequence of an unknown protein to the sequence of a protein whose structure is known, 2) generating the backbone conformation of loops and other regions which are not conserved between the two proteins, and 3) predicting the conformation of side chains, in both conserved and non-conserved regions of the structure. The current work deals only with the last part of this problem — modelling side chain conformation.

There is no generally accepted method for predicting side chain conformation, although several approaches have been developed. The methods may be divided into two broad categories: rule-based methods and conformation searching methods. Rule-based methods rely on the observation that topologically equivalent side chains in homologous structures generally adopt the same torsion angles (Summers *et al.*, 1987; Chothia *et al.*, 1989; Summers & Karplus, 1990). However, there are known cases in which identical side chains in homologous structures adopt different conformations (Summers *et al.*, 1987). Conformational searching approaches iteratively change the conformation of one

side chain at a time and score its conformation subject to a potential energy function (Snow, 1986; Bruccoleri, 1987; Novotny *et al.*, 1988). The method of Schiffer *et al.* (1990) has extended upon these methods by including a cycle of energy minimization prior to scoring. These algorithms are usually limited by the presence of multiple energy minima and so far solvation effects have not been taken into account (as a consequence, surface residues are generally modelled poorly). Furthermore, most conformational search methods developed to date are non-combinatorial in that they attempt to optimize only a single side chain at a time. Recent work by Lee and Subbiah (1991) has shown that the core residue conformation can be more accurately predicted using a Monte Carlo procedure that simultaneously optimizes the complete set of side chains for a protein.

Our approach is based on the observation by Ponder and Richards (1987) that side chain conformations can be modelled using a relatively small library of idealized 'rotamers.' Instead of considering the full conformational space theoretically accessible to a side chain, one of a small number of low energy conformations (typically 5-6 per residue type) can be used to describe the side chain. In trying to model the conformation of side chains, one need only specify which side chain rotamer will be observed, rather than independently predict the coordinates of each atom. Because proteins are well packed, the conformation adopted by one side chain can change the conformations of neighboring residues. This suggests that a successful side chain prediction algorithm must be multidimensional — simultaneously considering the combinatorial conformation space of several adjacent residues.

In previous studies, we have applied the rotamer representation of conformation space in combination with an approximate free energy force field to calculate the effects of mutagenesis on enzyme substrate specificity (Wilson *et al.*, 1991). Rotamers were used to model the three enzyme binding pocket residues and the P1 side chain of the substrate for mutants of the enzyme α-lytic protease. Surprisingly good agreement between calculated

and experimental binding energies was obtained for a diverse body of enzyme kinetic data ($r>0.85$, average energy error = 0.7 Kcal/mol for 42 mutant enzyme-substrate combinations). Crystallographic studies of the mutant proteases complexed with boronic acid peptide inhibitors (transition state-like analogs for the substrates) showed that the lowest energy combination of binding pocket side chain rotamers was generally the same as the set of rotamers observed in the x-ray structure. This observation suggested that the rotamer model could be used in a more general way to predict side chain conformation given a fixed protein backbone.

In developing our algorithm, we have deliberately avoided heuristic rules derived by examining pairs of known homologous proteins. While some success has been reported with such methods (Summers & Karplus, 1989), it remains unclear how generally and reproducibly one can apply these methods for true predictions. In the absence of many test cases, it is impossible to know which rules contribute to the accuracy of the side chain prediction and which are useful only for the test cases which have been studied.

As an alternative, we have used a molecular mechanics-based force field containing an additional solvation term to evaluate the suitability of each rotamer combination within a local site of ≈5 residues. The energy function, which has been parameterized to reproduce a large body of small molecule experimental data and enzyme kinetics measurements, provides all of the decision-making rules for selecting side chain conformations. As such, the test cases we have considered should mirror results obtained in true modelling suitations in which the correct structure is unknown.

Our algorithm for modelling side chains in an homologous structure presupposes that one knows the sequence alignment between two proteins. We start with a structure that has peptide backbone atoms included for only the conserved regions of the structure. Sites are defined iteratively throughout the protein and an attempt is made to optimize the local side chain packing within each site. All rotamers corresponding to the amino acids

within a site are combinatorially tested. An empirical free energy calculation is used to judge the quality of each possible rotamer combination and identify the lowest energy structure. Details of the method are described below.

We have verified this procedure using several test cases for which the structures of pairs of homologous proteins are known. Analysis of the structures before and after application of the above algorithm suggests that it is remarkably good at predicting side chain conformation, especially for pairs of proteins with high homology to one another.

Methods

(a) *algorithm outline*

Figure 9.1 outlines the key steps in the algorithm for modelling side chain conformations. After generating the starting model (lacking all side chain atoms), the coordinates for all rotamers appropriate to each amino acid type are calculated for all residues in the model. Rotamer coordinates are generated using the Ponder and Richards program for tertiary template calculations, PROPAK. Their standard rotamer library, containing 111 rotamers for all twenty amino acids, has been used for these calculations. All rotamers which do not make significant bad contacts with the main chain atoms are included in our combinatorial searching. After generating the rotamers, the following procedure is applied (see Figure 9.1):

1) A site center (one of the amino acids in the model) is chosen at random.

2) The N residues whose side chains are closest to the site center are identified.

3) For the $(N+1)$ residues in the site, all possible combinations of rotamers are tested, and for each rotamer combination, an approximate free energy is calculated.

4) After testing all combinations, the set of side chain rotamers which has the lowest calculated free energy is added to the model.

5) Steps 1-4 are repeated using a different randomly-chosen central amino acid until all residues in the model have been used as site centers.

6) Steps 1-5 are repeated until the predicted side chain conformations do not change from one cycle to the next.

For the test cases reported here, each site included 5 residues. Repeated cycles did not converge for all residues because local sites overlap and in several cases, the rotamer chosen by the algorithm depended on which other residues were being considered simultaneously. For all of the test cases, therefore, the procedure was terminated after having cycled through the entire protein three times. Increasing the size of the sites from 5 to 7 residues improved the results marginally but dramatically increased the computation time.

Details on the force field used to calculate the energy of each conformation are described in a previous paper (Wilson *et al.*, 1991). Since rotamers are used for the side chains, the internal geometry of the atoms is idealized and thus the energetic costs of altering bond lengths, bond angles, and dihedral angles can be neglected. The force field consequently has only two terms: 1) non-bonded interactions between atom pairs, and 2) the change in solvation energy upon exposing atoms to solvent. In our previous calculations of the effects of mutagenesis on enzyme substrate specificity, both the non-bonded and solvation terms were required to accurately reproduce experimentally-determined binding energies. The optimal weights for the two terms found for the α-lytic protease calculations have been used for these studies ($w_{Non\text{-}Bonded}=0.031$, $w_{Solvation}=1.98$). The non-bonded terms (including electrostatics, hydrogen-bonding, and van der Waals energies) are calculated using the AMBER molecular mechanics force field (with a

distance-dependent dielectric, $\varepsilon=r$, for the electrostatic term) (Weiner *et al.*, 1984). Non-bonded interactions greater than 100 Kcal/mole are truncated at this maximal value.

The effects of solvation are treated using a model similar to that of Eisenberg and McLachlan (1986). In their formalism, each atom is assigned an atomic solvation parameter (ASP); multiplying an atom's solvent accessible surface area by its ASP directly gives the solvation energy for that atom. We have used the ASP approach but rather than calculating the precise solvent accessible surface area for each atom (an extremely time-consuming computation), we use a grid method to estimate solvent accessibility. Each grid point on a 1.0-Å body-centered cubic lattice surrounding the macromolecule represents a pseudo-solvent molecule. For each conformation, a new list of allowed solvent positions is determined (solvent can be excluded from the grid by van der Waals contacts with non-solvent atoms or by a lack of hydrogen-bonding partners). The number of grid points surrounding an atom occupied by solvent is proportional to the total atomic accessible surface area and thus to the solvation energy for the atom. There is no statistically significant difference between the accuracy of the original Eisenberg-McLachlan method and our grid-based method, as judged by the ability to fit amino acid transfer free energies (Wilson *et al.*, 1991).

(b) *Building starting models*

To evaluate the performance of our side chain modelling algorithm, we constructed several 'homology-built' models using pairs of known protein structures. In each test, one protein was assumed to be unsolved (the 'unknown') and the other protein served as a starting point for the modelling (the 'template'). An effort was made to select pairs of structures with a wide range of sequence similarities. The sequence identities between the pairs we tested (listed in Table 9.1) varied from 30% to 100%. The models were made by substituting the amino acid side chains of the unknown structure onto the backbone of the template structure at all unambiguously equivalent positions. To determine which positions

were structurally equivalent, the crystal structures were first superimposed manually using computer graphics, then refined automatically by finding all pairs of residues within 2.5 Å of each other. This alignment was then corrected by hand: in some cases where a loop in one structure had a similar conformation but was displaced by more than 2.5 Å from the analogous loop in the other structure, residues were still assigned to be equivalent. Loops having very different conformations and insertions were omitted from the models.

(c) *Model Evaluation*

To evaluate the resulting side chain conformations, the backbones of the true and model structures were superimposed by a least-squares fit method. The r.m.s. deviations of the side chains and backbones of every residue in the model and true structures were then computed. We have tabulated both a side chain-average r.m.s. deviation (averaging the r.m.s. deviation calculated for each residue), and an overall r.m.s. deviation (summed over all side chain atoms). The average r.m.s. deviation (used by Novotny *et al.*, 1988) is generally smaller than the overall r.m.s. deviation since side chains with fewer atoms tend to be more accurately modelled and their contribution is more highly weighted in the average deviation than in the overall deviation.

We also generated a best possible structure using the template structure backbone and the rotamer library. At each position, the rotamer having the lowest r.m.s. deviation from the side chain in the true structure was selected. This 'lowest coordinate error' structure is the best result we could hope to obtain with the algorithm since we allow only idealized rotamers for the side chains. To score our model, we determined the fraction of the side chains that had been assigned to the same rotamer as that in the lowest error structure.

To determine whether our procedure actually improves the accuracy of the model, we also generated structures which had the most common rotamer assigned to each residue.

The r.m.s. deviation of this unrefined 'first guess' model from the true structure gives a baseline measure against which other models can be compared.

Results

(a) *Idealized test case with α-lytic protease*

Before applying the algorithm to homology model building, we first tested the rotamer approximation and the force field to determine how accurately we could identify the correct side chain rotamers under the best possible conditions. In the first test case, the side chains from the structure of α-lytic protease were removed one at a time and then built back on to the protein backbone. To best isolate possible sources of error, the following changes were made in the modelling procedure. 1) For each side chain, the rotamer in the library closest to the observed side chain (in terms of r.m.s. deviation) was replaced by the true side chain conformation. 2) All symmetry-related atoms within 10Å of any protein atom were included in the energy calculations (since crystal contacts may determine the conformation of surface side chains) and the two crystallographically-determined sulfate ions were explicitly included in the calculation. 3) Instead of combinatorially searching all rotamers at sites throughout the protein, a single residue was searched at a time and after identifying the lowest-energy conformation, the original true side chain was temporarily added back to the protein. At the completion of the calculation, the predicted rotamers were then built onto the structure at each position. The only possible sources of error in this test case are the crystallographic coordinates for α-lytic protease and the force field used to estimate the free energy.

Of 142 residues with more than one rotamer (*i.e.* not glycine or alanine residues), the crystallographic side chain rotamer was identified as the lowest energy rotamer in 126

cases (89% correct). The overall r.m.s. deviation between the lowest energy structure and the crystal structure was 0.59 Å (side chain atoms only, see Table 9.2). While the majority of side chains are modelled correctly, there are certainly a significant number of errors in this best case model.

Assuming that the incorrect predictions result from errors in the force field used to evaluate them, we hoped to better understand these problems by analyzing the characteristics of the poorly placed side chains. Of the 16 residues which were not correctly predicted, 12 were exposed and 4 were buried. The bias towards exposed residues is not unexpected since there are strong packing constraints on buried residues which do not exist at the surface (especially when only a single side chain is varied at a time). Surprisingly, all four incorrectly predicted buried residues were polar amino acids, including three serines and one asparagine (Val 40 was the only hydrophic amino acid among the 16 incorrect residues). For most of the incorrect side chains, both non-bonded and solvation terms are lower for the incorrect rotamer than for the correct rotamer, suggesting that adjustments in the weighting between these two terms are unlikely to improve the prediction.

The incorrect residues include an unusually high number of serines (6) and asparagines (5), perhaps indicating that uncharged hydrogen bonds may be treated improperly by the force field. In their work on side chain modelling, Summers and Karplus (1989) note that the energy surface for the rotation of serine hydroxyls is often characterized by multiple minima with energies close to that of the global minimum. This observation would explain the tendency for serine side chains to be poorly predicted. Often, the incorrectly predicted side chains form a set of hydrogen bonds that differs from that actually found in the crystal structure. In addition, for several residues both the predicted and the observed side chains are pointing out into solution and there appears to be no obvious reason why either conformation is preferred over the other.

It is possible that incorrectly placed side chains are indicative of errors in the crystal structure rather than of errors in the force field. For four residues (Asn-62, Asn-118, Asn-162, and Ser-189), the calculated energy difference between the crystal structure side chain rotamer and the lowest energy rotamer is more than 10 Kcal/mole — it seems unlikely that force field errors alone could make the crystal-structure side chain conformation appear so unfavorable. The original α-lytic protease structure (as with most crystal structures) was refined without considering the interactions of protein hydrogen atoms (Fujinaga *et al.*, 1985). Bad contacts involving side chain hydrogen atoms are found for the four incorrectly-predicted crystal structure side chains with large energy errors. Analysis of a $2|F_o| - |F_c|$ omit map (F_c calculated using the crystal structure with the incorrectly predicted side chains omitted) shows, however, that these four high energy residues (and all other misplaced side chains) fit the electron density quite well. In one case (Asn 62), hydrogen bad contacts could be relieved by a 180° flip of χ_2, switching the positions of OD1 and ND2 but not changing the fit to the electron density. For most other residues, the crystal structure side chain rotamer provides an unambigously better fit to the electron density than the predicted rotamer. Interestingly, if the four high energy crystal structure side chains are substituted by the appropriate equivalent library rotamer with ideal geometry, all but one residue (Asn-162) are now correctly predicted. This observation strongly suggests that the crystal structure side chains are modelled with the correct rotamer but that small adjustments in atomic position are required to properly model hydrogen-bonding interactions.

The conformation of many solvent-accessible residues may be determined in the crystal by contacts with symmetry-related molecules. In a true homology-building exercise, it will be impossible to model the crystal contacts. To test this possibility, we carried out the calculation again, this time leaving out symmetry-related molecules and bound counter-ions. The effect appeared to be negligible as the overall r.m.s. deviation

increased only slightly to 0.62 Å with no net change in the number of incorrectly modeled side chains (Table 9.2). In all subsequent tests, symmetry related atoms and bound sulfates have been ignored.

In the first test case we replaced the rotamer in the library with the lowest r.m.s. deviation by the true side chain at each site. This was done to remove the possibility that errors in the results were due to the assumption of idealized rotamer geometry. In the third test case, we repeated the test using only the standard library rotamers to determine whether the rotamer approximation would severely hinder the modeling procedure when true side chains are not known. Using the library rotamers, the overall r.m.s. deviation for side chain atoms increased from 0.62 Å to 1.21 Å (116 of 142 side chains (82%) were modeled correctly). The lowest error rotamer structure (calculated using the library rotamers) has an overall r.m.s. deviation of 0.71 Å; the predicted structure is thus only 0.5 Å worse than the best structure possible given the constraint to ideal rotamer geometry. Amino acids with long side chains (lysine, arginine, glutamine) account for more than two thirds of the residues which were initially predicted but are incorrectly placed after reverting to the standard rotamer library. This bias is not surprising since the rotamer approximation should be worst for those amino acids with many torsion angles.

In the above simulations, each residue test was done in the context of an otherwise correct structure. To test the ability of the algorithm to converge without the correct neighboring side chains, the above test was repeated with a starting structure that was completely stripped of side chains. Using the standard algorithm (with sites of five residues, cycling through the sequence three times, adding the lowest energy rotamer combination in each case), the number of correct side chains plateaued at 111 residues (versus 116 residues in the previous test). This result indicates that the combinatorial conformation search converges well in the absence of a starting bias towards the correct structure.

The general conclusions of the α-lytic protease test cases (summarized in Table 9.2) are as follows. 1) The force field is able to correctly predict almost 90% of the observed side chain conformations, with the incorrectly predicted side chains lying largely at the surface and including a disproportionately high number of serines and asparagines. 2) Symmetry-related atoms and bound counter-ions do not significantly affect the ability to predict side chain conformation. 3) Using side chain rotamers rather than the true side chains prevents the correct prediction of ≈7% of the residues. 4) By combinatorially searching local sites throughout the protein, it is possible to accurately predict most side chains without a starting bias to the correct structure. With no errors in the protein backbone but no starting information about side chain conformations, the overall r.m.s. deviation for α-lytic protease side chains falls to 1.31 Å after refinement. Similarly good results were obtained for a number of other proteins (Table 9.3), indicating the general utility of the algorithm.

(b) *Homology Modeling*

With the accuracy of the force field and the rotamer assumption well tested, we have proceeded to use the algorithm to predict side chain conformations for pairs of homologous proteins. These homology modeling tests differ from the α-lytic protease test cases in two respects: 1) the backbone used to calculate side chain rotamers is not exactly correct (since the template structure backbone differs slightly from the true structure backbone), and 2) the template peptide backbone is punctuated by gaps where non-homologous regions have not been modelled. Our results are summarized in Table 9.4. In every case, there is a significant improvement in the accuracy of the model following application of the algorithm. The ability to correctly predict side chain conformation decreases as the deviation between the model backbone and true backbone increases. The improvement, as measured by r.m.s. deviation to the true structure or by the fraction of

correctly predicted side chains, drops approximately linearly with decreasing sequence identity (Figure 9.2).

As with the α-lytic crystal structure test case, solvent-accessible residues are signficantly harder to predict than buried residues. Figure 9.3 shows the predicted structure of hen egg-white lysozyme, with both hydrophobic core residues and some surface residues. While aromatic residues making up the core are all accurately positioned, exposed residues are often incorrect. The fraction of buried or solvent accessible residues that are correctly placed for each test case are listed in Table 9.4. Increased errors at the surface are likely to be due to a number of phenomena including greater allowed conformational space (since there are fewer restricting adjacent residues), increased crystallographic errors in the surface residues, and errors in the force field that effect electrostatic interactions more than van der Waals interactions (given that hydrophilic residues predominate at the surface). Previous analysis of protein crystal structures has shown that surface residues have systematically higher temperature factors than buried residues (Alber *et al.*,1987), indicating that their side chain atoms are less well fixed in an energy minimum. This observation suggests that the energy differences between alternate conformations may be smaller at the surface and that slight errors in the force field should affect surface residues more than buried ones.

By comparing the predictions made for the ALP→ALP test and the SGB→ALP test, we can quantify the errors introduced by using the wrong backbone to predict side chain conformations. Using the standard iterative procedure to place α–lytic protease side chains on the backbone of α–lytic protease, 78% of side chains (111/142) are correctly predicted. This fraction drops to 63% (76 / 121) when the backbone of *S. griseus* protease B is used instead. The average r.m.s. deviation of side chains also increases in going from the α-lytic protease backbone (0.73 Å) to the *S. griseus* protease B backbone (1.88 Å). This increase is higher than that observed for the backbone atoms (0.00 Å for ALP, 0.79 Å

for SGB), suggesting that errors in the backbone positions adversely affect the choice of side chain rotamer, beyond simply displacing the side chain away from the correct position.

Figure 9.4 shows a representative case in which deviations in the backbone between a pair of homologous structures directly lead to an incorrect side chain choice. The backbone atoms of the neighboring residues isoleucine 105 and tyrosine 237 in *S. griseus* protease B deviate by only ≈0.4 Å relative to their equivalents in α-lytic protease (tryptophan 105 and tyrosine 238). These relatively small deviations alter the direction of the $C_\alpha \rightarrow C_\beta$ vectors sufficiently to cause a significant change in the positions of the calculated rotamers. The r.m.s. deviation of the lowest error rotamers at these positions from the true side chains rises from 0.2 Å (using the α-lytic backbone) to 1.4 Å (using the *S. griseus* protease B backbone). More importantly, the SGB→ALP lowest error rotamer structure has several bad van der Waals contacts between tryptophan 105 and tyrosine 238, causing this combination of rotamers to be ignored during the rotamer search. For instance, the separation between NE1 of tryptophan 105 and CG of tyrosine 238 drops from the close distance of 3.02 Å in the ALP→ALP lowest error rotamer structure to the bad contact distance of 2.43 Å in the SGB→ALP structure (Figure 9.4). Whereas the lowest error rotamers for residues 105 and 238 are identified as the lowest energy combination in the ALP→ALP test, an alternate set of rotamers is chosen for the SGB→ALP case. The same is true of Arginine 48A which lies adjacent to this pair of residues. While it is correctly placed for the ALP→ALP test case, the incorrectly positioned tyrosine 238 in the SGB→ALP test forces this arginine into an incorrect position.

Discussion

This work has shown that a combinatorial rotamer search directed by an approximate free energy calculation can be used to predict side chain conformation in a homology modelling test. The fraction of properly placed side chains is a function of the similarity between the pair of homologous structures, dropping from ≈80% in the case of 100% identity, to ≈60% for those tests with lower homology. By using rigid rotamers to coarsely sample conformation space and a grid approach to calculate solvent accessibility, the complete combinatorial search can be carried out extremely quickly. Starting with the backbone alone, the prediction of side chain conformation for a 200-residue protein can be completed in less than 5 hours of VAX 8650 CPU time. Our algorithm provides a significant improvement, both in terms of accuracy and speed, over energy-based side chain modelling algorithms that have been previously reported (Bruccoleri *et al.*, 1987; Schiffer *et al.*, 1990). The reasons for this improvement will now be considered.

The CONGEN program of Bruccoleri *et al.* (1987) uses a grid search over main chain and side chain torsion angles to model both loops and side chain conformation. The conformation space of each added side chain is searched individually and evaluated using the CHARMM force field. This molecular mechanics force field includes terms for covalently-linked atom pairs (bond-stretching, bond angle bending, torsion angle rotation) and for non-bonded pairs (van der Waals' forces, electrostatics, and hydrogen-bonding). After evaluating all staggered conformations, the lowest energy conformation is saved. This method can replace side chains onto a structure with the correct backbone with an r.m.s. deviation of ≈2.5 Å (averaged over side chains, not including C_β atoms).

While conceptually similar to the approach we have described, there are several major differences between the two methods. In contrast to our program, CONGEN includes bonded-energies but ignores solvation effects in evaluating side chains. Since our

approach uses rotamers with idealized internal geometry, there is no need to consider the bonded terms. Work by Bruccelori and others, however, has shown that the side chain rotamers preferred by a molecular mechanics force field lacking solvation terms are biased towards those that have relatively unfavorable solvation energy (Novotny *et al.*, 1988; Schiffer *et al.*, 1990). By not taking into account solvation effects during the rotamer search, therefore, the CONGEN approach tends to incorrectly predict the conformation of polar surface side chains (Novotny *et al.*, 1988).

A second difference between the two methods is in their approach to sampling conformation space. Whereas the CONGEN program can search an arbitrary number of rotamers for a single side chain, our algorithm combinatorially tests a limited number of rotamers at each site for a cluster of adjacent residues. By simultaneously varying several side chain conformations, energetic barriers to co-operative rearrangements can be surmounted. In support of this, if our algorithm is applied using sites containing a single residue rather than five adjacent residues, an additional ≈10% of the side chains are not correctly predicted after the first cycle (data not shown).

Schiffer *et al.* (1990) describe a method for constructing side chains that is closely-related to the CONGEN approach. In this algorithm, staggered side chain conformations are evaluated using the AMBER force field (Weiner *et al.*, 1984). A zone surrounding each targetted residue is subject to energy minimization to improve the packing around the altered side chain. The final minimized energy for each side chain orientation is used to determine which rotamer is adopted at each site. As with the CONGEN algorithm, this approach does not take into account solvation effects and does not combinatorially test adjacent side chains. It does, however, have the significant advantage of allowing side chains to deviate from their initial idealized rotamer geometry. In cases in which slight bad contacts exist between the library rotamers (*e.g.* Figure 9.4), energy minimization should allow the contacting atoms to relax and thereby yield a more realistic energy estimate.

Because several hundred thousand cycles of energy minimization must be done to complete a single cycle of side chain optimization, this approach is extremely computer-intensive. It has currently been applied to only a subset of the residues in the one test case which has been reported (bovine→rat trypsin). As presently implemented, this method seems promising but it may require a significant increase in computer-speed to be generally practicable.

Recent work by Lee and Subbiah (1991) has applied the technique of simulated annealing to the problem of predicting side chain conformation. Their algorithm uses a force field containing only torsional and van der Waals terms to evaluate any given combination of side chains. Discreet steps of 10° in the side chain torsion angles are taken to sample different conformations. Rather than defining a limited site within the protein, the complete set of protein side chains is simultaneously optimized, using an annealing Monte Carlo procedure to coarsely sample a wide section of conformation space. This procedure has been applied to a number of test cases in which the correct backbone is used but all side chain atoms are initially deleted. While good results are obtained for buried residues (average r.m.s. deviation = 1.25 Å), errors in the surface residues are significantly higher, causing the average overall r.m.s. deviation of side chain atoms to rise to 1.97 Å. Results obtained using our algorithm (Table 9.3) indicate that the use of a more complete force field (including electrostatic and solvation terms) significantly improves upon their predictions. Whereas buried side chains are predicted with approximately equal accuracy by either the Lee-Subbiah algorithm or ours, our overall r.m.s. deviations are ≈25% lower; indicating a significant improvement for surface residues. *A priori*, one might have expected the Lee-Subbiah algorithm to more accurately predict buried side chain conformation since their method allows a much finer conformation search (36 conformers / torsion angle versus 5-6 rotamers / side chain with our method). The fact that buried

residues are predicted equally well by both methods indicates that the assumption of rotamer geometry for side chains is usually a very good approximation.

We have demonstrated how small errors in the backbone coordinates can force the incorrect choice of side chain rotamers. By restricting the conformation search using a rigid backbone, it is currently impossible to overcome these problems. It may be possible to significantly improve upon our results by carrying out energy minimization or molecular dynamics on the full structure between the cycles of side chain optimization. If the side chains are correctly modelled onto the template backbone, relaxation of the structure using energy minimization or molecular dynamics should allow backbone atoms to move towards their true coordinates, generating the 'unknown' backbone. As the side chain conformations gradually improve between cycles, the ability to minimize to the correct backbone should also increase. The potential benefits of this approach may be limited by the accuracy of the force field used to direct the energy minimization. In most simulations starting from a crystal structure, energy minimization causes r.m.s. shifts of ≈0.5—1.0 Å. However, for side chain modelling in cases of low homology this level of error is likely to be small enough to allow an improvement over the current fixed-backbone procedure.

Acknowledgements

Funding for this research was provided by the Howard Hughes Medical Institute. CW was supported by a Fannie and John Hertz Foundation fellowship in applied physics. LMG was supported by a predoctoral fellowship from the National Science Foundation.

Table 9.1: Test cases for side chain conformation optimization

model (template→ unknown)	N_{unk}	N_{tmpl}	N_{mod}	overall similarity (%)	similarity of modeled regions (%)	resol. unk (Å)	resol. tmpl (Å)	backbone rms.dev. (Å)
ALP→ALP	198	198	198	100.0	100.0	1.7	1.7	0.00
LZ1→LYZ	129	130	129	60.2	60.5	1.5	2.0	0.61
LBP→LIV	344	346	344	79.1	79.4	2.4	2.4	0.69
SGB→ALP	198	185	168	33.4	37.5	1.8	1.7	0.79
PTN→SGT	223	223	204	29.7	35.3	1.7	1.7	0.98

The 'unknown' proteins were modeled from the 'template' structures. N_{unk}: number of residues in the unknown structure. N_{tmpl}: number of residues in the template structure. N_{mod}: number of superimposable residues in the unknown and template structures which were modeled in the predicted structure. Overall similarity: percent sequence identity between the unknown and template structure determined using the sequence alignment method of Smith and Smith (1989). Similarity of modeled regions: percent sequence similarity for superimposable residues between unknown and template structures. Resol. unk: crystallographic resolution of the unknown structure. Resol. tmpl: crystallographic resolution of the template structure. Backbone r.m.s.d.: root-mean-square deviation of backbone coordinates between the unknown and template structures for the residues modeled.

Structures used (Brookhaven PDB entry names in parentheses): ALP: α-lytic protease (2ALP) (Fujinaga *et al.*, 1985); SGB: protease B from *S. griseus* (3SGB) (Read *et al.*, 1983); SGT: *S. griseus* trypsin (1SGT) (Read *et al.*, 1984); PTN: bovine trypsin (3PTN) (Walter *et al.*, 1982); LYZ: hen egg white lysozyme (6lyz) (Diamond *et al.*, 1974);

LZ1: human lysozyme (1LZ1) (Artymiuk *et al.*, 1981); LIV: leucine/isoleucine/valine binding protein (2LIV) (Sack *et al.*, 1989a); LBP: leucine binding protein (2LBP) (Sack *et al.*, 1989b).

Table 9.2: Results of the α-lytic protease test cases

Starting with the xtal structure side chains?	Symmetry-related molecules and counter-ions?	Replace library rotamer with xtal structure ?	Size of site / temporarily force correct rotamer?	average r.m.s. deviation (Å)	overall r.m.s. deviation (Å)	fraction correct (correct/total)
yes	yes	yes	1 / yes	0.26±0.58	0.59	0.89 (126 / 142)
yes	no	yes	1 / yes	0.27±0.61	0.62	0.89 (126 / 142)
yes	no	no	1 / yes	0.68±0.85	1.21	0.82 (116 / 142)
no	no	no	5 / no	0.73±0.91	1.31	0.76 (111 / 142)

Results for modelling the side chains of α-lytic protease using the true backbone are shown. Average r.m.s. deviation: average root-mean-square deviation of non-alanine side chains of the predicted structure from the true structure (unweighted by the number of atoms in each side chain). Overall r.m.s.deviation: root-mean-square deviation of all side chain atoms.

Table 9.3: Predicting side chain conformation using the correct peptide backbone

Protein	Lowest error rotamer structure r.m.s. deviation	Final overall r.m.s. deviation	Final buried residue r.m.s. deviation (% correct)	Final accessible residue r.m.s. deviation (% correct)
2ALP	0.71	1.31	0.93 (88)	1.56 (67)
1CRN	0.38	1.32	0.82 (88)	1.41 (66)
5PTI	0.67	1.67	1.76 (69)	1.65 (58)
1CTF	0.63	1.49	1.06 (75)	1.60 (34)

Side chains were initially deleted from all proteins and then built on using the standard algorithm described in the text. Solvent accesibility was calculated using the method of Lee and Richards as implemented in the program ACCESS by Hundschumather and Richards. Residues considered buried have less than 20% of their accessible surface aresase exposed (relative to an extended tripeptide model). Labels correspond to the Protein Data Bank file containing the starting coordinates: 2ALP- α-lytic protease (Fujinaga *et al.*, 1985), 1CRN- crambin (Hendrickson and Teeter, 1981), 5PTI- bovine pancreatic trypsin inhibitor (Walter and Huber, 1983), and 1CTF- L7/L12 50S ribosomal protein (Leijonmarck and Liljas, 1987).

Table 9.4: Homology modelling results

Structure	first guess average rmsd	first guess overall rmsd	predicted average rmsd	predicted overall rmsd	L.E. average rmsd	L.E. overall rmsd	fraction correct buried	fraction correct exposed	fraction correct total
ALP→ALP	1.69±1.31	2.48	0.73±0.91	1.31	0.39±0.44	0.71	0.88 (58/66)	0.67 (52/78)	0.78
LZ1→LYZ	1.99±1.21	2.58	1.44±1.02	1.90	0.91±0.49	1.06	0.78 (32/41)	0.53 (34/64)	0.63
LBP→LIV	2.05±1.16	2.49	1.55±1.05	1.96	1.10±0.56	1.25	0.78 (96/123)	0.43 (61/143)	0.59
SGB→ALP	2.29±1.47	3.03	1.88±1.50	2.66	1.30±1.07	1.73	0.70 (37/53)	0.57 (39/68)	0.63
PTN→SGT	2.40±1.60	3.17	1.95±1.60	2.68	1.44±1.27	1.92	0.81 (54/67)	0.46 (38/83)	0.61

Results for homology modelling test cases. Abbreviations for the proteins are the same as in Table I. L.E. - lowest error structure.

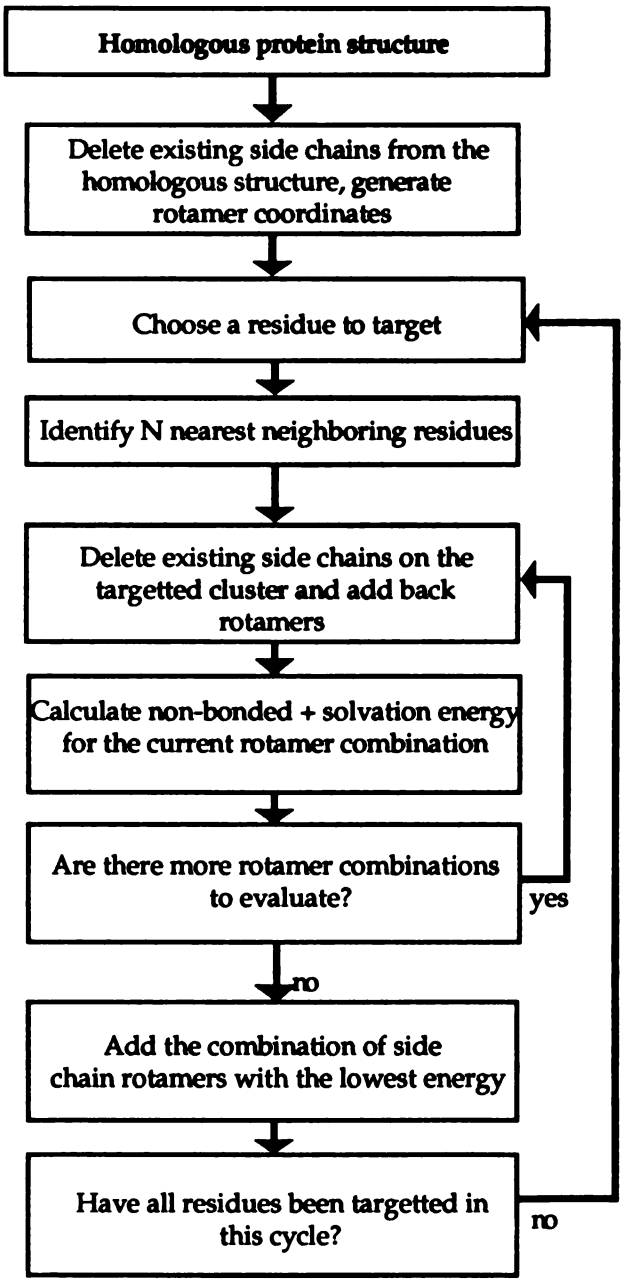


FIGURE 9.1. Algorithm for predicting side chain conformation. Each step is descibed in detail in the text.

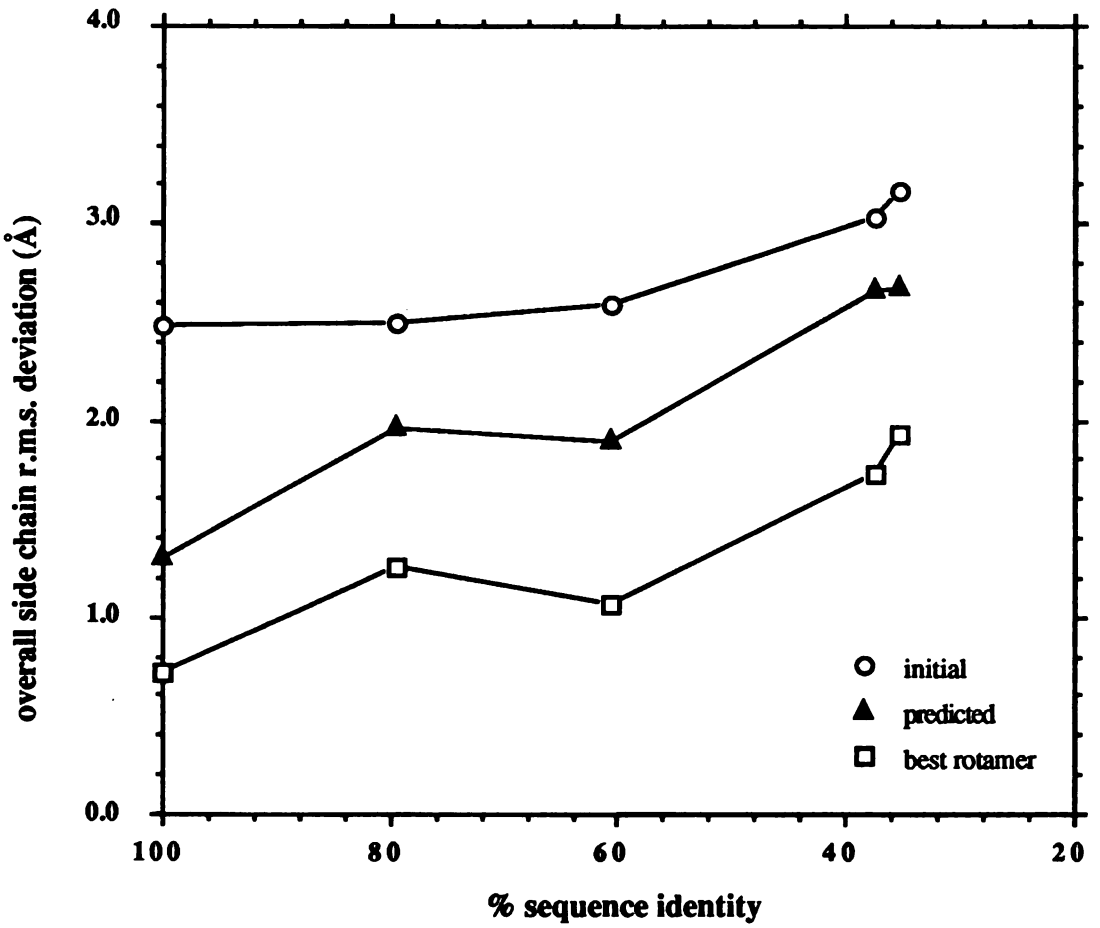


FIGURE 9.2. Accuracy of the side chain prediction as a function of percent homology. The overall side chain r.m.s. deviation is shown as a function of the percent homology between the 'unknown' and 'predicted' structures for the initial, predicted, and 'best rotamer' (lowest r.m.s. deviation) models.

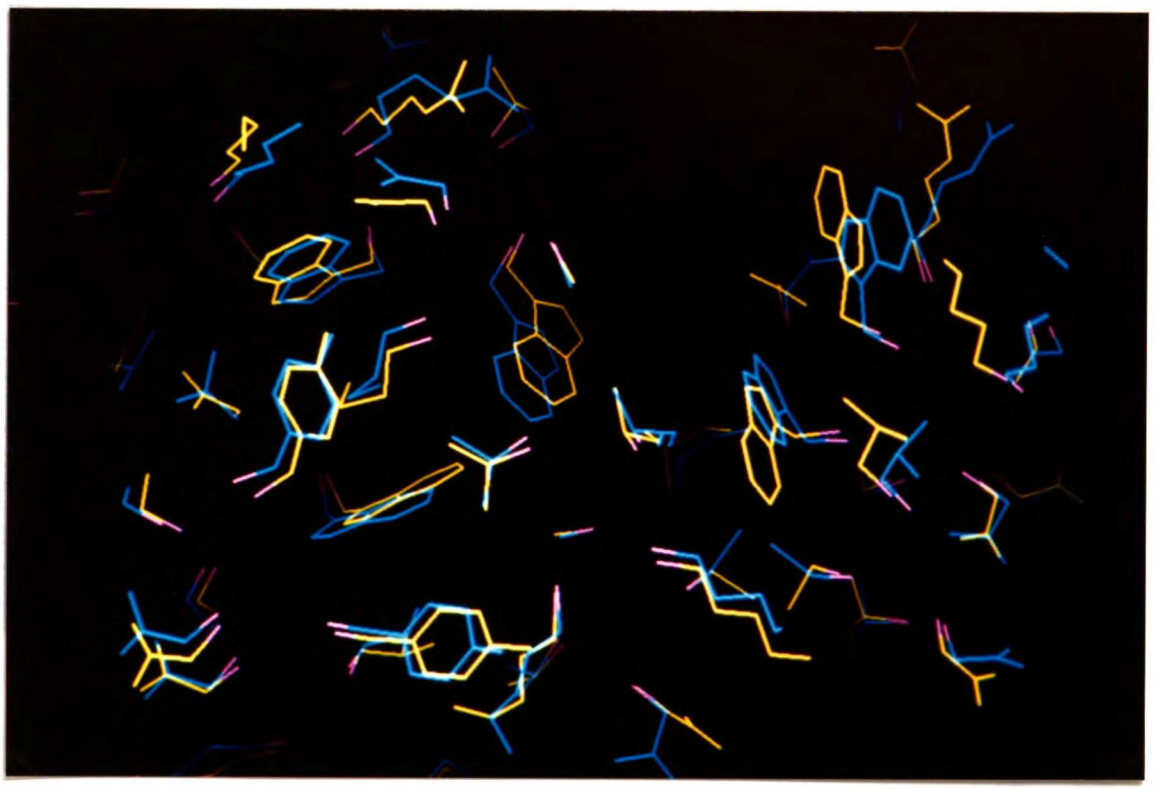


FIGURE 9.3. Comparison between the predicted and observed hen eggwhite lysozyme structures. Side chain and C_α atoms are shown for the predicted (blue) and true (yellow) structures (C_α atoms colored magenta). Residues in the hydrophobic core lie on the left-hand side while those on the right are generally somewhat solvent accessible.

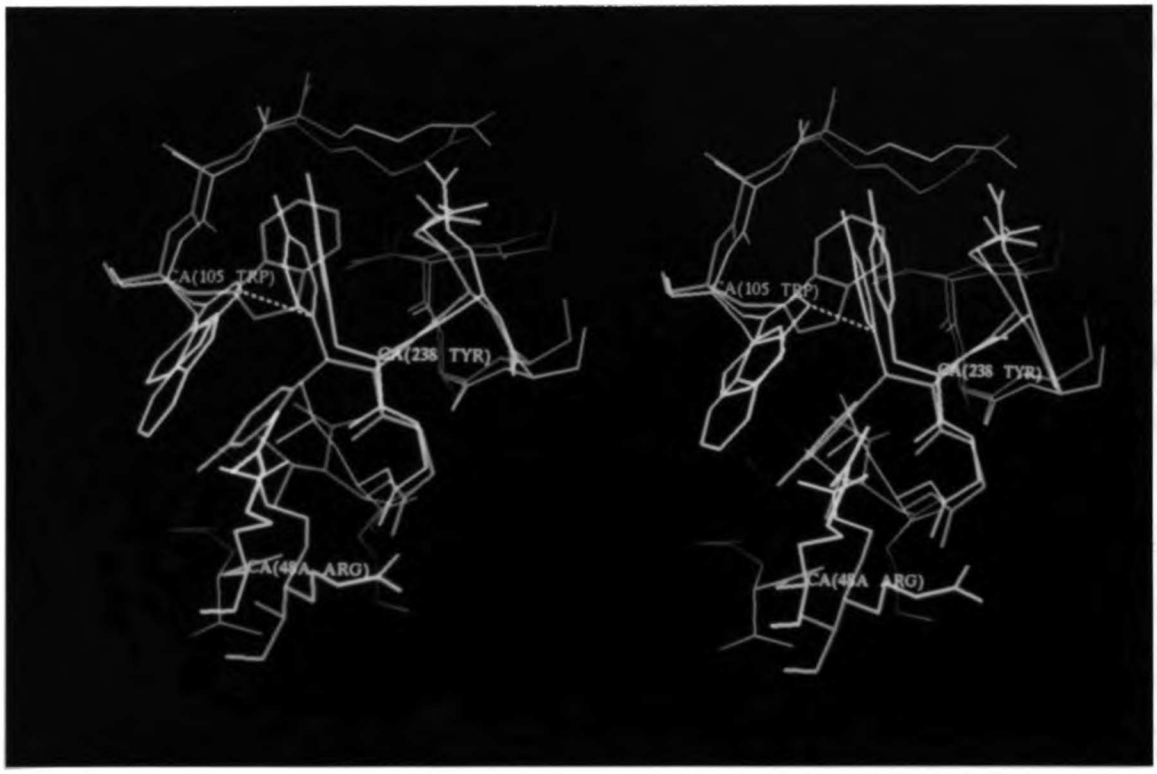


FIGURE 9.4. Errors in the *S. griseus* protease B → α-lytic protease prediction. The true (yellow) and predicted (blue) structures for α-lytic protease are shown (using *S. griseus* protease B as a backbone template). The 'best' rotamers (those with the lowest r.m.s. deviation to the true structure) are shown in magenta. Several bad contacts between the best rotamers for Trp 105 and Tyr 238 (*e.g.* NE1-105 - CG-238 distance = 2.43 Å, dotted line) force an alternate set of rotamers to be chosen as the lowest energy conformation. The misplaced tyrosine 238 ring subsequently forces Arg 48A to adopt an incorrect conformation. All three residues are correctly positioned when using the true α-lytic protease backbone to generate the side chain rotamers (not shown).

References

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. (1987). Crystallographic Databases -- Information Content, Software Systems, Scientific Applications. Bonn/Cambridge/Chester, Data Commission of the International Union of Crystallography.

Alber, T., Sun, D.P., Nye, J.A., Muchmore, D.C., & Matthews, B.W. (1987). Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry.* **26**, 3754-3758.

Artymiuk, P.J. & Blake, C.C.F. (1981). Refinement Of Human Lysozyme At 1.5 Angstroms Resolution. Analysis Of Non-bonded And Hydrogen-bond Interactions. *J.Mol. Biol.,* **152**, 737-762.

Bash, P.A., Singh, U.C., Langridge, R. & Kollman, P.A. (1987). Free Energy Calculations by Computer Simulation. *Science,* **236**, 564-568.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures . *J. Mol. Biol.,* **112 (3)**, 535-542.

Bruccoleri, R.E. & Karplus, M. (1987). Prediction of folding of short polypeptide segments by uniform conformational sampling. *Biopolymers,* **26**, 137-168.

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., *et al.* (1989). Conformations of immunoglobulin hypervariable regions. *Nature,* **342**, 877-883.

Diamond, R. (1974). Real-space Refinement Of The Structure Of Hen Egg-white Lysozyme. *J. Mol. Biol.,* **82**, 371-391.

Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199-203.

Fujinaga, M., Delbaere, L.T.J., Brayer, G.D. & James, M.N.G. (1985). Refined Structure Of Alpha-lytic Protease At 1.7 Angstroms Resolution. Analysis Of Hydrogen Bonding And Solvent Structure. *J. Mol. Biol.*, **184**, 479-502.

Hendrickson, W.A. & Teeter, M.M. (1981). Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature*, **290**, 107-113.

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

Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal domain of the ribosomal protein L7-L12 from Escherichia coli at 1.7 angstroms. *J. Mol. Biol.*, **195**, 555-579.

Novotny, J., Rashin, A.A. & Bruccoleri, R.E. (1988). Criteria That Discriminate Between Native Proteins and Incorrectly Folded Models. *Proteins: Struct. Func. Genet.*, **4**, 19-30.

Ponder, J.A. & Richards, F.M. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequence for different structural classes. *J. Mol. Biol.*, **193**, 775-791.

Read, R.J., Fujinaga, M., Sielecki, A.R. & James, M.N.G. (1983). Structure of the complex of *Streptomyces griseus* protease B and the third domain of the turkey ovomucoid inhibitor at 1.8 angstroms resolution. *Biochemistry*, **22**, 4420-4433.

Read, R.J. & James, M.N.G. (1984). Critical comparison of comparative model building of *Streptomyces griseus* trypsin. *Biochemistry*, **23**, 6570-6575.

Sack, J.S., Saper, M.A. & Quiocho, F.A. (1989). Periplasmic binding protein structure and function. Refined X-ray structures of the leucine/isoleucine/valine-binding protein and its complex with leucine. *J. Mol. Biol.*, **206**, 171-191.

Sack, J.S., Trakhanov, S.D., Tsigannik, I.H. & Quiocho, F.A. (1989). Structure of the L-leucine-binding protein refined at 2.4 Angstroms resolution and comparison with the leu/ile/val-binding protein. *J. Mol. Biol.*, **206**, 193-207.

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

Snow, M.E. & Amzel, L.M. (1986). Calculating the three-dimensional changes in protein structure due to amino acid substitutions: The variable region of immunoglobulins. *Proteins: Struct., Func., Genet.*, **1**, 267-279.

Summers, N.L., Carlson, W.D. & Karplus, M. (1987). Analysis of side-chain orientations in homologous proteins. *J. Mol. Biol.*, **196**, 175-198.

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

Walter, J. & Huber, R. (1983) Pancreatic trypsin inhibitor. A new crystal form and its analysis. *J. Mol. Biol.*, **167**, 911-917.

Walter, J., Steigemann, W., Singh, T.P., Bartunik, H., Bode, W. & Huber, R. (1982). On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography. *Acta Crystallogr.,Sect.B*, **38**, 1462-1472.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, **106**, 765-784.

FOR REFERENCE

NOT TO BE TAKEN FROM THE ROOM

BRODART CAT. NO. 23 012 PRINTED IN U.S.A.