Title
Packing and protein structure

Author
Gregoret, Lydia Maria

Publication Date
1991

Peer reviewed|Thesis/dissertation

Packing and Protein Structure

by

Lydia Maria Gregoret

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Chemstry

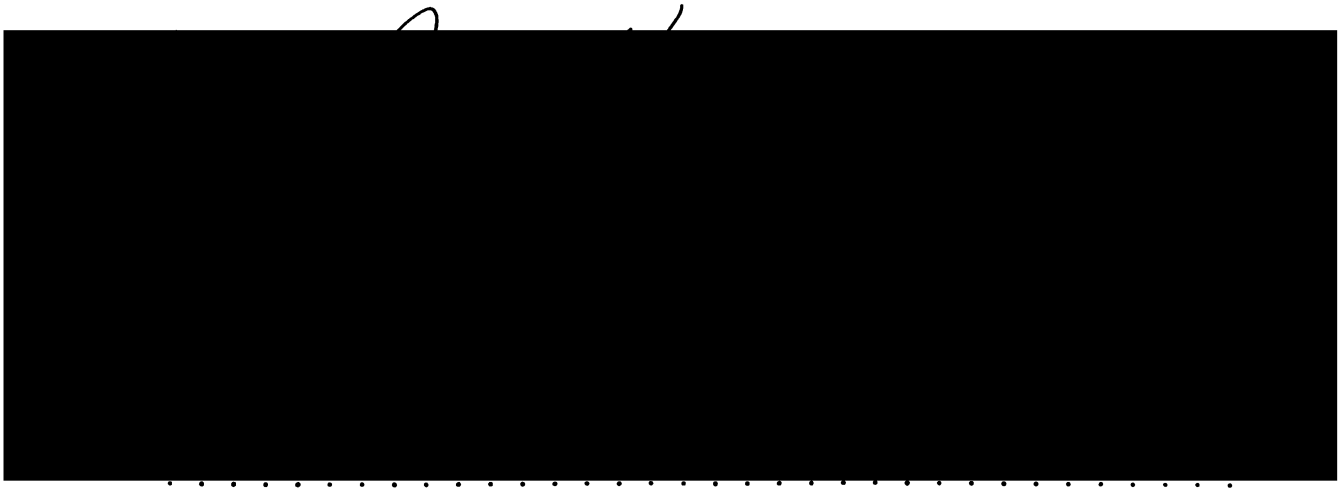
in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco



Date . . . . . University Librarian

Degree Conferred: 3/25/91

To Chuck Wilson and Olga Petrenko

Acknowledgments

First and foremost, I would like to thank my mentor, Fred Cohen. Fred is an extraordinary teacher. I hope that some day I too will be able to intellectually challenge my own students while treating them as fairly and kindly as he does. I also thank the other members of my thesis committee, Tack Kuntz and Robert Fletterick, not only for reading my thesis but especially for numerous scientific discussions over the last five years, and for their valuable advice and continuing encouragement.

Open discussions and collaborations make UCSF a special place. The cercarial elastase project (Chapter 4) brought the unlikely combination of protein folders and parasitologists together. I learned a lot about enzymology and schistosomiasis from Jim McKerrow and his students, Payman Amiri and Johnny Railley. Thanks also to Stephen Rader for his enthusiastic contribution to the sulfur hydrogen bond project (Chapter 5).

I want thank all of the members of the Cohen Group, particularly Scott Presnell for being a good sport and conversationalist about both proteins and cars, and also for keeping our group computers superbly maintained. I am forever grateful to Don Kneller for taking the time to teach me to program while he was working on *his* thesis (and I forgive him for the occasional eavesdropping over the room divider.)

I had a great time in graduate school thanks to all of my friends at UCSF, particularly Brian Shoichet and Dave Yee, with whom Chuck and I partook in numerous hill/wine trips, late afternoon Baker Beach barbeques and orals clubs at the Tassajara bakery where we frightened the other customers with our loud discussions about hydrogen *bonds*. And I can't forget to thank Karen Schultz, my friend of sixteen years, for having the good sense to go to Berkeley for graduate school so we wouldn't have to spend so much money on phone bills.

I would not have received a Ph.D. without the support of my family. I cannot ever thank my parents, Oleg and Maya, enough for insisting on giving me the best possible education, both by initiating creative activities at home and by sending me to excellent schools. I dedicate my thesis to my husband, Chuck Wilson, and to my grandmother, Olga Petrenko. It was my grandmother, herself a science teacher, who taught me about the joys of science. When I was young, she gave me numerous books about nature and bought me a microscope. Together we performed many experiments in her kitchen and garden. Today she takes avid interest in Chuck's and my research projects while continuing her own research in the garden. I thank her for being my first science teacher and for being a fan. She is a most caring and remarkable person.

Chuck is my best friend. I couldn't imagine sharing the experience of graduate school with anyone else. With no one else have I had more discussions about science. He is a great person to bounce ideas off of, no matter how stupid the ideas. As with my grandmother, Chuck and I have also performed many experiments at home, but this time on cars rather than vegetables. At work, we have successfully collaborated on a side chain modeling project (Chapter 4). I am very much looking forward to sharing the rest of our lives and careers.

Packing and Protein Structure
by
Lydia Maria Gregoret

Abstract

This thesis addresses the protein folding problem from a computational perspective. A central theme to the work presented here is amino acid packing in proteins. A method for evaluating packing in model-built, or predicted protein structures is described in Chapter 2. The effects of compactness on protein folding are studied in Chapter 3. Chapter 4 is a description of a method of protein structure prediction known as "modeling by homology," along with an application of this technique to an enzyme involved in schistosomal infection, and a method for modeling the conformations of side chains in homology-built structures. Chapter 5 describes the frequency and significance of hydrogen bonds involving sulfur atoms in proteins, and Chapter 6 describes unusual packing of proteases.

Table of Contents

List of Tables

List of Figures

Chapter 1.

Introduction

Proteins are crucial to every cellular function. They are involved in the catalysis of chemical reactions, the transport of molecules across cell membranes and through the blood, cell - cell communication, cell motility and muscle contraction, and cell division and growth. The three-dimensional structures of proteins are therefore interesting because they give us an understanding of the mechanisms of these actions at a molecular level.

Protein structures are currently determined using x-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). These methods are time-consuming and depend on technical variables such as whether a protein will crystallize (in the case of x-ray crystallography) or whether a protein is soluble in water at high concentration (in the case of NMR). The structures of only several hundred proteins have been determined using these techniques. In contrast, the number of protein sequences determined using DNA cloning and sequencing methods is in the tens of thousands. It would be extremely useful to have an efficient way to translate these sequences into structures.

Thus, the protein folding problem may be stated as follows: how does the amino acid sequence of a protein (its *primary structure*) determine its three dimensional (or *tertiary*) structure? It was first shown by Christian Anfinsen and co-workers (Anfinsen *et al.*, 1961) that the small protein, ribonuclease A, could refold *in vitro* into an active catalyst from a large collection of denatured conformations. It is hypothesized that the active, folded conformation is found by the polypeptide chain because it is at a global free energy minimum (Kauzmann, 1959; Dill, 1990). Though more than thirty years have passed since Anfinsen's fundamental experiments, no one has yet discovered how the three-dimensional structure of a protein is encrypted in its sequence.

This thesis addresses the protein folding problem from a computational perspective. A central theme to the work presented here is amino acid packing in proteins. Globular proteins, those which are soluble in water, are observed to be very well-packed (Richards, 1977). Protein interiors do not contain holes or large, water-filled

cavities, and the volumes occupied by amino acids in a protein are not unlike the volumes of the amino acids in pure crystals (Chothia, 1975). It is remarkable that proteins are so efficiently packed. Efficient packing, however, may be inevitable. A protein is probably forced to be compact by the hydrophobic effect which seeks to minimize exposed, non-polar surface area in water (Chan & Dill, 1990; Lau & Dill, 1990). There is evidence that even when the volume of several nearby, internal residues is changed dramatically, the protein accommodates by rearranging slightly, but still retaining the same overall fold (Lesk & Chothia, 1980; Lim & Sauer, 1989; Lim & Sauer, submitted)

A method for the evaluation of packing in proteins and the effects of compactness on protein folding are presented in this thesis. The thesis is organized as follows. In Chapter 2, a new method for evaluating the quality of predicted protein structures on the basis of amino acid packing and amino acid distribution is described. Evaluating packing in model-built protein structures can give a measure of model quality and can serve to help sort between alternative predictions. The method uses a simplified representation of the polypeptide chain where amino acids are modeled as spheres (Gregoret & Cohen, 1990).

Chapter 3 is a study of the significance of compactness in driving secondary structure formation. It has long been assumed that the hydrogen bonding capability of the polypeptide backbone is responsible for the preponderance of α-helices and β-sheets in proteins. Recent studies by Chan and Dill (1990) suggest that compactness, arising from the hydrophobic effect which seeks to minimize exposed, non-polar surface area in water, may account for a significant fraction of the secondary structure we see in proteins. By exhaustively searching conformation space on a cubic lattice, these workers found that in maximally compact structures, fifty percent of residues, on average, were in either a helical or a sheet-like conformation. In order to determine if compactness alone can explain the secondary structure content of real proteins, I have tested the Chan-Dill hypothesis *off* the lattice by generating more realistic structures using a self-avoiding random walk. The results of this work suggest that compactness

does indeed promote the formation of secondary structure. Without the constraint of a lattice, however, most of the structure which forms is helical. Sheet formation is rare, suggesting that other forces, perhaps hydrogen bonds, may be necessary to bring together strands which are sequentially distant from one another along the polypeptide chain.

Chapter 4 focuses on a method of protein structure prediction known as modeling by homology. This is the most successful method of structure prediction because the structure of the protein of interest is modeled based on a related protein whose three-dimensional structure has been determined experimentally. Chapter 4 is divided into three parts. Part A is an introduction to modeling by homology. Part B describes the modeling of an elastinolytic protease implicated in the invasion of human skin by the parasite *Schistosoma mansoni* and subsequent inhibitor design based on this model. This work was done in collaboration with Dr. James McKerrow's group at UCSF (Cohen *et al.*, submitted). The schistosomal elastase has sequence similarity to the trypsin-like serine proteases, and the model was built using pancreatic elastase as a scaffold. Part C of Chapter 4 describes a technique for predicting the conformations of side chains in homology-built models (Wilson *et al.*, in preparation). This method uses the Ponder and Richards (1987) library of side chain rotamers. The conformations of nearby side chains are combinatorially changed to optimize amino acid packing and side chain - side chain interactions subject to a force field which contains both non-bonded and solvation terms.

Chapters 5 and 6 are observations about protein structures. Chapter 5 is a study of the frequency of occurence of hydrogen bonds involving sulfur atoms in proteins (Gregoret *et al.*, 1991). Sulfur's ability to hydrogen bond has been ignored in previous reviews of hydrogen bonding in proteins. With the popularity of site-directed mutagenesis the propensities of all amino acid side chains should be understood fully.

Chapter 6 is a short study of unusual packing in proteases. A study of the distribution of alpha-carbon - alpha-carbon distances in a large data set of proteins showed that the number of short distances in proteases is unusually high. Proteases also

seem to have slightly lower surface area : volume ratios than other proteins. These observations suggest that proteases may have convergently evolved to be more well-packed or to have less-convoluted surfaces in order to be more resistant to proteolysis.

Three of the five chapters described above (numbers 2, 3, and 5) have been published previously. Thus, each chapter is self-contained, with introduction, methods, results, discussion, and conclusions sections. References pertaining to each chapter are found at the end of that chapter.

References for Chapter 1

Anfinsen, C.B., Haber, E., Sela, M. & White, F.H. (1961). The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *Proc. Natl. Acad. Sci., USA.* **47**, 1309.

Chan, H.S. & Dill, K.A. (1990). Origins of Structure in Globular Proteins. *Proc. Natl. Acad. Sci. USA.* **87**, 6388-6392.

Chothia, C. (1975). Structural Invariants in Protein Folding. *Nature (London).* **254**, 304-308.

Cohen, F.E., Gregoret, L.M., Amiri, P.A., Aldape, K., Railey, J. & McKerrow, J.H. (submitted). Arresting Tissue Invasion of a Parasite with Synthetic Protease Inhibitors Chosen by Computer Modeling. *Biochemistry.*

Dill, K.A. (1990). Dominant Forces in Protein Folding. *Biochemistry.* **29**, 7133-7155.

Gregoret, L.M., Rader, S.D., Fletterick, R.J. & Cohen, F.E. (1991). Hydrogen Bonds Involving Sulfur Atoms in Proteins. *Proteins: Struct., Func., Genet.* **9**, 99-107.

Kauzmann, W. (1959). Some Factors in the Interpretation of Protein Denaturation. *Adv. Protein Chem.* **14**, 1-62.

Lau, K.F. & Dill, K.A. (1990). Theory for Protein Mutability and Biogenesis. *Proc. Natl. Acad. Sci., USA.* **87**, 638-642.

Lesk, A.M. & Chothia, C. (1980). How Different Amino Acid Sequences Determine Similar Protein Structures: The Structure and Evolutionary Dynamics of the Globins. *J. Mol. Biol.* **136**, 225-270.

Lim, W.A. & Sauer, R.T. (1989). Alternative Packing Arrangements in the Hydrophobic Core of λ repressor. *Nature.* **339**, 31-36.

Lim, W.A. & Sauer, R.T. (submitted). The Role of Internal Packing Interactions in Determining the Structure and Stability of a Protein. *J. Mol. Biol.*

Ponder, J.A. & Richards, F.M. (1987). Tertiary Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes. *J. Mol. Biol.* **193**, 775-791.

Richards, F.M. (1977). Areas, Volumes, Packing and Protein Structure. *Ann. Rev. Biophys. Bioeng.* **6**, 151-176.

Wilson, C., Gregoret, L.M. & Agard, D.A. (in preparation). Modeling Side Chain Conformations for Homologous Proteins Using an Energy-Based Rotamer Search.

Chapter 2.

A Novel Method for the Rapid Evaluation of Packing in Protein Structures†

† This chapter has been published in the Journal of Molecular Biology, volume 211, pages 959-974, 1990.

Introduction

It would be of tremendous benefit to the general understanding of biological processes to be able to predict the three-dimensional structure of a protein from its amino acid sequence. The current methods for the determination of high-resolution structures (x-ray crystallography, nuclear magnetic resonance) have fallen significantly behind the rate of sequence publication. Here we present a method for evaluating *predicted* protein structures for correctness to assist in model building.

It is often possible to predict the overall fold of a novel sequence if the structure of a closely related protein (> 35% sequence similarity) has already been solved (Greer, 1981; Greer, 1985; Strynadka & James, 1988; Taylor, 1988). If the structure of *more* than one homologous protein is known, the quality of the model generally improves (Blundell *et al.*, 1987; Taylor, 1986; Zvelebil *et al.*, 1987). It is difficult, however, to detect errors in homology-built models. Energy minimization does not conveniently lend itself to the analysis of regional errors, particularly the misalignment of secondary structure elements. Loops and side chains are difficult to model because they have many degrees of conformational freedom. Also, identical residues in identical backbone positions in homologous structures do not always adopt the same side chain dihedral angles (Summers *et al.*, 1987).

"De novo" modeling, that is, predicting the structure of a novel sequence that bears no resemblance to any known structures, is much more challenging. Additional difficulties arise because many alternative solutions can appear feasible, and preliminary models tend to be much more speculative than models built by homology. Most approaches to *de novo* structure prediction apply a hierarchical scheme of first predicting secondary structure (Chou & Fasman, 1974; Cohen *et al.*, 1983; Cohen *et al.*, 1986b; Garnier *et al.*, 1978; Lim, 1974a; Lim, 1974b), and then predicting the tertiary fold (Cohen *et al.*, 1983; Cohen *et al.*, 1986a; Cohen & Sternberg, 1980b; Cohen *et al.*, 1980; Cohen *et al.*, 1982). To predict the tertiary fold, we use a combinatorial

method to generate every possible assembly of secondary structural elements. This generally results in about one hundred structures which satisfy the topological and spatial constraints of all residues being connected in a fixed linear sequence. These resulting structures are rather crude, in that only the backbone is modeled. Because of the lack of detail in these structures, evaluating them by conventional energy calculations has been ineffective (Cohen, Sternberg & Robson, unpublished data). A few close non-bonded interactions can overwhelm any correct features of the model.

An important consideration in all model-building strategies is that of locating errors in the model. In order to construct a better model, one must be able to determine the errors in the current model. Novotny and co-workers (1984; 1988) purposely created incorrect model structures in order to help answer this question. Using two small proteins, hemerythrin (an all-alpha four-helical bundle), and immunoglobulin kappa chain (a beta sheet sandwich), they modeled each sequence onto the other's backbone and minimized the energy of the model structures. Their initial conclusion was that it was not possible to distinguish the "misfolded" models from the correct models on an energetic basis alone. Upon closer examination, however, it could be seen that the misfolded structures did not conform to certain known features of globular proteins: Hydrophobic residues were located on the exterior of the protein and unpaired charges on the interior. It was postulated that these alternative folded forms would not be stable in aqueous solution.

Richards (1974; 1977) notes that proteins elegantly solve the packing problems posed by a compact heterologous chain. Here we present a method for quickly evaluating packing density. In addition, we derive a measure of the quality of pairwise amino acid contacts. In our model, amino acids are represented as spheres (the aromatic amino acids are modeled as either two or three spheres.) Our representation of the protein is based loosely on the centroid-based model of Levitt and Warshel (1976; 1975) used for folding simulations.

Using our method, and with the assistance of computer graphics, it is possible to qualitatively determine which regions of the model protein are poorly packed in terms of side chain steric overlap. We are able to differentiate between alternative combinatorially-produced structures, and to select a subset of "best" structures using a residue-residue interaction potential. Our new method for evaluating packing will be a valuable adjunct to model building and combinatorial structure prediction schemes.

Methods

1. The Sphere Growth Method

The polypeptide chain is represented as a series of spheres. With the exception of aromatic amino acids, histidine, phenylalanine, tyrosine and tryptophan, each amino acid is modeled as one sphere located at the center of mass of all side chain atoms plus the alpha carbon. The aromatic amino acids histidine, phenylalanine and tyrosine are modeled as two spheres: one at the β-carbon and the other at the centroid of the aromatic ring atoms. Tryptophan is modeled as three spheres: one at the β-carbon, another at the center of the five-membered ring, and a third at the center of the six-membered ring. The coordinates of the backbone atoms, N, Cα, C, and O are retained for viewing purposes, but ignored in the actual packing calculation.

To evaluate packing density and pairwise interactions, each amino acid-sphere is assigned an "ideal" radius (r_i^0) according to its amino acid type (see Table II.1). The "actual" radius of each amino acid (r_i) is initially set equal to zero. Then, all of the radii in the protein are increased at a rate proportional to r_i^0. Once the sum of the *actual* radii of two residues i and j, $(r_i + r_j)$, is equal to the separation between their centroids d_{ij}, their sizes are fixed. Figure 2.1 depicts three snapshots of the

Table II.1

residue name	sample size	ideal radius (Å)	ideal radius (Å) (ring)	ideal radius (Å) (ring 2)	volume (Å^3)
ALA	1388	2.25 ± 0.23	-- --	-- --	47.7
ARG	564	2.53 ± 0.50	-- --	-- --	67.8
ASN	712	2.42 ± 0.36	-- --	-- --	59.4
ASP	945	2.38 ± 0.37	-- --	-- --	56.5
CYS	131	2.25 ± 0.21	-- --	-- --	47.7
CYX	250	2.43 ± 0.26	-- --	-- --	60.1
GLN	582	2.51 ± 0.38	-- --	-- --	66.2
GLU	838	2.54 ± 0.42	-- --	-- --	68.6
GLY	1539	2.02 ± 0.25	-- --	-- --	34.5
HIS	379	2.35 ± 0.33	2.39 ± 0.50	-- --	97.8
ILE	842	2.53 ± 0.25	-- --	-- --	67.8
LEU	1208	2.54 ± 0.25	-- --	-- --	68.6
LYS	998	2.65 ± 0.45	-- --	-- --	78.0
MET	262	2.45 ± 0.28	-- --	-- --	61.6
PHE	615	2.27 ± 0.29	2.23 ± 0.30	-- --	87.6
PRO	734	2.28 ± 0.28	-- --	-- --	49.6
SER	1324	2.29 ± 0.29	-- --	-- --	50.3
THR	1071	2.40 ± 0.28	-- --	-- --	57.9
TRP	237	2.26 ± 0.27	2.22 ± 0.33	2.12 ± 0.28	110.7
TYR	570	2.29 ± 0.30	2.15 ± 0.35	-- --	87.9
VAL	1253	2.50 ± 0.26	-- --	-- --	65.4

Ideal sphere sizes for the 21 amino acids with standard deviations (across a dataset of 72 structures) and the resulting residue volumes. ("CYX" is half-cystine.) The aromatic amino acids, histidine (HIS), phenylalanine (PHE) and tyrosine (TYR) are represented by two spheres, "beta" and "ring." The beta sphere is located at the β-carbon of the side chain, and the ring sphere is located at the centroid of the aromatic ring atoms. Tryptophan (TRP) is represented by three spheres: one at the β-carbon, another at the centroid of the five-membered ring and a third at the centroid of the six-membered ring (designated here as "ring2") The volume of each of these residues is the volume enclosed by their two or three intersecting spheres.

sphere-growing process. In this Figure, r_i is increased in finite increments. Initially, we used a step size of 0.01 x r^0. As the step size is made infinitely small, the analytical solution to this problem emerges. The current version of our program uses the analytical approach in which we determine first the *effective* distance between all pairs of residues, $d_{ij} / (r_i^0 + r_j^0)$, then identify the shortest effective distances to establish the unique order in which "growing" residues will bump into each other. This order depends solely on the positions of the centroids in space and their ideal radii. Once a residue's growth is terminated, all effective distances to that residue are adjusted to reflect the inability of this residue to grow any larger. As we are interested in non-bonded contacts, the radius of the residue CYX (half-cystine) is allowed to interpenetrate its crosslink partner CYX. Similarly, the individual spheres comprising the aromatic amino acids are allowed to interpenetrate each other.

To derive an r^0 for each amino acid or centroid type, the sphere growth calculation was performed on a data set of 72 protein structures from the Brookhaven Protein Data Bank (Abola *et al.*, 1987; Bernstein *et al.*, 1977) (see Table II.2). The data set was selected from structures which were non-degenerate and for which complete atomic detail was available at an atomic resolution of 2.5Å or less. Initially, the ideal radii of 3.0Å were used for every amino acid sphere. At the conclusion of the calculation, statistics were tabulated for each amino acid type to determine the average *actual* sphere size attained for that type of residue over the data set of proteins. These average sizes were designated as the ideal sizes and the calculation was repeated on the data set of proteins. Again, at the conclusion of the calculation, average sizes were tabulated. After five iterations, the average sizes and their associated standard deviations were observed to converge (Table II.1). The resulting ideal radii did not depend on the starting ideal radii (all 3.0 Å): the same radii resulted when the initial ideal radii were all set to either 5.0Å or 1.0Å. Nor did the resulting ideal radii depend on the algorithm used (stepwise or analytical).

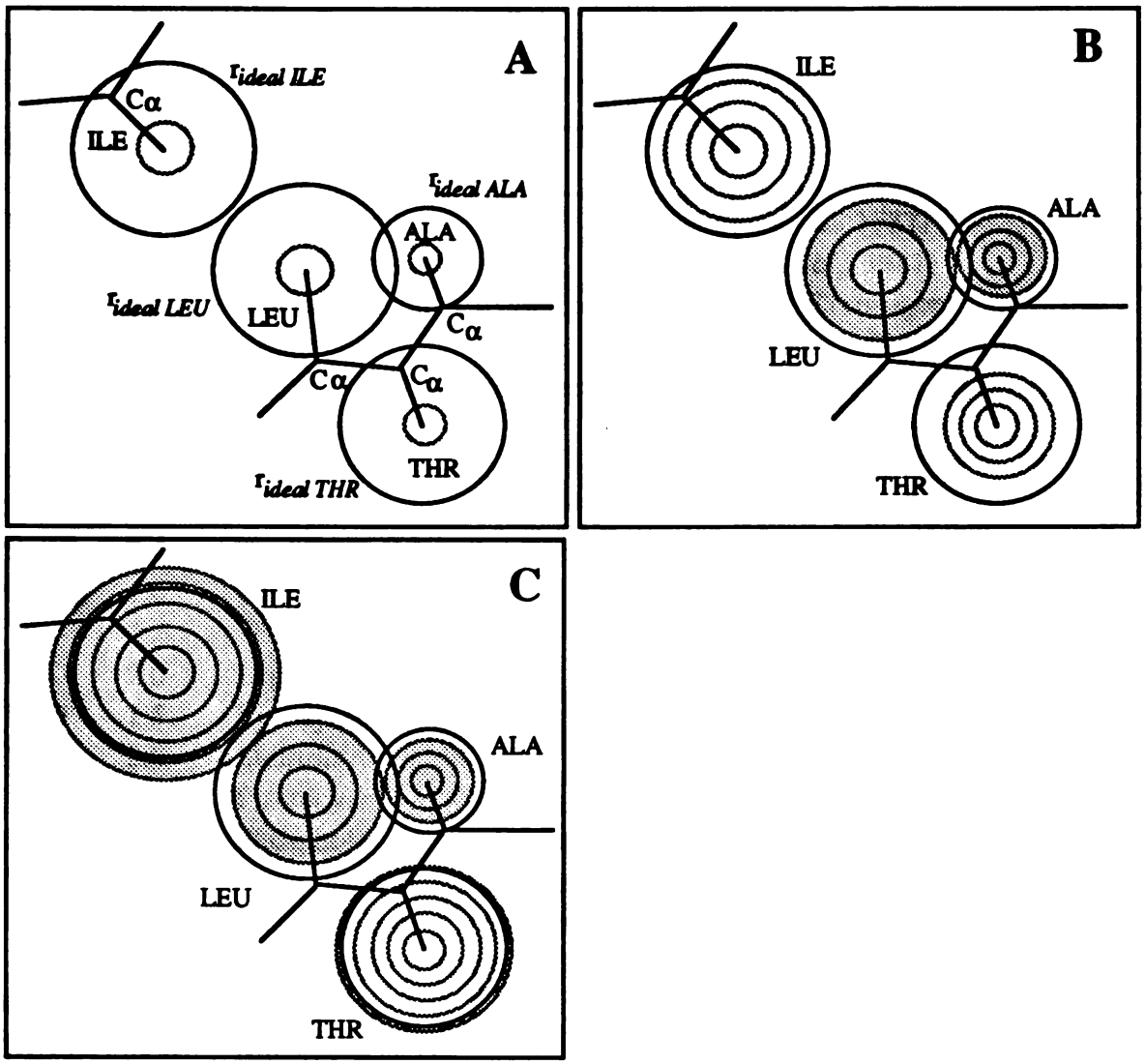


Figure 2.1 Frames **A**, **B** and **C** show the stages of the packing calculation. The centroids and alpha-carbons of four residues in a hypothetical protein are shown. The solid circles represent the "ideal" radii, r^o, for the amino acids depicted. The relative sizes of the amino acids are exaggerated. In frame **A**, the radius of each residue has been incremented by one step (lighter circles.) Note that the step size is proportional to the ideal radius for each amino acid type. For the sake of clarity, the step sizes have been exaggerated. Frame **B** shows of the calculation after three steps. The sum of the radii of alanine and leucine, $r_{ALA} + r_{LEU}$, is now equal to the distance between their centroids, $d_{ALA\text{-}LEU}$, so the radii of these residues are fixed. Alanine and leucine are colored grey to indicate that they have reached their final size. These residues both achieved approximately 80% of their ideal size before their growth was terminated ($r_{ALA}/r^o_{ALA} = r_{LEU}/r^o_{LEU} = 0.8$). Frame **C** shows the status of the calculation after five steps. Now, isoleucine is also colored grey because $r_{ILE} + r_{LEU} = d_{ILE\text{-}LEU}$. Isoleucine exceeded its ideal radius by approximately 20% before its radius was fixed ($r_{ILE}/r^o_{ILE} = 1.2$). Threonine, after five steps, still has not come in contact with any other residues and will continue having its radius incremented until it contacts one of the other residues. Two pairwise interactions have occurred so far in this scenario: ALA-LEU and LEU-ILE.

Table II.2
Data set of protein structures used to derive ideal radii for amino acid-centroids

PDB entry	resolution (Å)	protein name	reference
1ACX	2.0	Actinoxanthin	(Pletnev *et al.*, 1982)
1BP2	1.7	Phospholipase A2 (cow)	(Dijkstra *et al.*, 1981)
1CPV	1.85	Calcium-binding Parvalbumin B	(Moews & Kretsinger, 1975)
1CRN	1.5	Crambin	(Hendrickson & Teeter, 1981)
1CTF	1.7	L7/L12 50S Ribosomal Protein	(Leijonmarck & Liljas, 1987)
1ECO	1.4	Erythrocruorin	(Steigemann & Weber, 1979)
1FB4	1.9	Immunoglobulin Fab	(Marquart *et al.*, 1980)
1FDX	2.0	Ferredoxin	(Adman *et al.*, 1976)
1FX1	2.0	Flavodoxin (D. vulgaris)	(Watenpaugh *et al.*, 1980)
1GCR	1.6	Gamma-II Crystallin	(Slingsby *et al.*, unpublished data submitted to the PDB)
1LH1	2.0	Leghemoglobin	(Arutyunyan *et al.*, 1980)
1LZ1	1.5	Lysozyme (human)	(Artymiuk & Blake, 1981)
1MLT	2.0	Melittin	(Terwilliger & Eisenberg, 1982)
1PCY	1.6	Plastocyanin	(Guss & Freeman, 1983)
1PP2	2.5	Phospholipase A2 (rattlesnake)	(Brunie *et al.*, 1985)
1PPT	1.37	Avian Pancreatic Polypeptide	(Blundell *et al.*, 1981)
1REI	2.0	Bence-Jones Immunoglobulin REI Variable Portion	(Epp *et al.*, 1975)
1RHD	2.5	Rhodanese	(Ploegman *et al.*, 1978)
1RN3	1.45	Ribonuclease A	(Borkakoti *et al.*, 1982)
1RNT	1.9	Ribonuclease T1	(Arni *et al.*, 1988)
1SBT	2.5	Subtilisin BPN	(Alden *et al.*, 1971)
1TIM	2.5	Triose Phosphate Isomerase	(Banner *et al.*, 1976)
1TON	1.8	Tonin	(Fujinaga & James, 1987)
1UBQ	1.8	Ubiquitin	(Vijay-Kumar *et al.*, 1987)
2ABX	2.5	Alpha-Bungarotoxin	(Love & Stroud, 1986)
2ACT	1.7	Actinidin	(Baker & Dodson, 1980)
2ALP	1.7	Alpha-lytic Protease	(Fujinaga *et al.*, 1985)
2APP	1.8	Penicillopepsin	(James & Sielecki, 1983)
2AZA	1.8	Azurin	(Baker, 1988)
2CAB	2.0	Carbonic Anhydrase Form B	(Kannan *et al.*, unpublished data submitted to the PDB)
2CCY	1.67	Cytochrome c'	(Finzel *et al.*, 1985)
2CNA	2.0	Concanavalin A	(Reeke *et al.*, 1975)

Table II.2 continued

PDB entry	resolution (Å)	protein name	reference
2CPP	1.63	Cytochrome P450cam	(Poulos *et al.*, 1987)
2CYP	1.7	Cytochrome c Peroxidase	(Finzel *et al.*, 1984)
2HHB	1.74	Hemoglobin (Deoxy)	(Fermi *et al.*, 1984)
2LZM	1.7	Lysozyme (Bacteriophage T4)	(Weaver & Matthews, 1987)
2MDH	2.5	Malate Dehydrogenase	(Birktoft & Banaszak, 1983)
2OVO	1.5	Ovomucoid Third Domain	(Bode *et al.*, 1985)
2PAB	1.8	Prealbumin	(Blake *et al.*, 1978)
2PKA	2.05	Kallikrein A	(Bode, Epp *et al.*, 1985)
2PRK	1.5	Proteinase K	(Betzel *et al.*, 1988)
2SGA	1.5	Proteinase A (S. griseus)	(Moult *et al.*, 1985)
2SNS	1.5	Staphylococcal Nuclease	(Legg, 1977)
2SOD	2.0	Superoxide Dismutase	(Tainer *et al.*, 1982)
3ADK	2.1	Adenylate Kinase	(Dreusicke *et al.*, 1988)
3APR	1.8	Rhizopuspepsin	(Suguna *et al.*, 1987)
3CLN	2.2	Calmodulin	(Babu *et al.*, 1988)
3EBX	1.4	Erabutoxin B	(Smith *et al.*, 1988)
3EST	1.65	Porcine Elastase	(Meyer *et al.*, 1988)
3FAB	2.0	Lambda Immunoglobulin Fab	(Saul *et al.*, 1978)
3GAP	2.5	Catabolite Gene Activator Protein	(Weber & Steitz, 1987)
3GRS	1.54	Glutathione Reductase	(Karplus & Schulz, 1987)
3ICB	2.3	Calcium-binding Protein	(Szebenyi & Moffat, 1986)
3PGK	2.5	Phosphoglycerate Kinase	(Bryant *et al.*, 1974)
3RP2	1.9	Rat Mast Cell Protease II	(Remington *et al.*, 1988)
3SGB	1.8	Proteinase B (S. Griseus)	(Read *et al.*, 1983)
3TLN	1.6	Thermolysin	(Holmes & Matthews, 1982)
4ADH	2.4	Apo-liver Alcohol Dehydrogenase	(Eklund *et al.*, 1976)
4APE	2.1	Acid Proteinase Endothiapepsin	(Pearl *et al.*, unpublished data submitted to the PDB)
4CHA	1.68	Alpha-chymotrypsin	(Tsukada & Blow, 1985)
4CYT	1.5	Cytochrome c	(Takano & Dickerson, 1980)
4DFR	1.7	Dihydrofolate Reductase	(Bolin *et al.*, 1982)
4FXN	1.8	Flavodoxin	(Smith *et al.*, 1977)
4LDH	2.0	Lactate Dehydrogenase	(White *et al.*, 1976)
4PTI	1.5	Trypsin Inhibitor	(Marquart *et al.*, 1983)

Table II.2 continued

PDB entry	resolution (Å)	protein name	reference
4PTP	1.34	Beta Trypsin	(Chambers *et al.*, unpublished data submitted to the PDB)
4RXN	1.2	Rubredoxin	(Watenpaugh, Sieker *et al.*, 1980)
4TNC	2.0	Troponin C	(Satyshur *et al.*, 1988)
5CPA	1.54	Carboxypeptidase A	(Rees *et al.*, 1983)
6LYZ	2.0	Lysozyme (hen egg white)	(Diamond, 1974)
7CAT	2.5	Catalase	(Fita & Rossmann, 1985)
9PAP	1.65	Papain	(Kamphuis *et al.*, 1984)

2. Derivation of Pair Potentials

To compute a pair potential matrix describing the tendencies of particular amino acids to contact each other, all residue-residue interactions, or "bumps" which occurred during the sphere growing procedure were recorded. (See Figure 2.1.) Then, in order to increase the number of contacts per residue, the sphere-growing procedure was repeated on the data set, this time allowing residues to grow *through* their nearest neighbor until they made contact with their *next* nearest neighbor. Combining the pairwise interactions resulting during these first and second rounds of sphere growth, the number of interactions of a given type (e.g. ALA-LEU) was divided by the number of interactions expected given a random distribution of residues throughout the protein. The natural logarithm of this ratio multiplied by -1 gives the "pair potential" for that type of interaction. For example, if n_{AL} ALA-LEU interactions occurred in the data set and there were N_A interacting alanines, N_L interacting leucines and N_{tot} interacting residues total (thus $N_{tot}/2$ total interactions), the pair potential (PP[A][L]) would be calculated as follows:

$$PP[A][L] = -\ln\left(\frac{\frac{2n_{AL}}{N_{tot}}}{\frac{N_A}{(N_{tot})}\frac{N_L}{(N_{tot}-1)}+\frac{N_L}{(N_{tot})}\frac{N_A}{(N_{tot}-1)}}\right) = -\ln\left(\frac{n_{AL}(N_{tot}-1)}{N_A \cdot N_L}\right)$$

The sum in the denominator of the middle expression reflects the fact that ALA-LEU interactions and LEU-ALA interactions are equivalent. This ratio may be thought of as a pseudo-equilibrium constant. Thus, by analogy to free energy, a pair potential less than zero is favorable. The pair potential matrix is shown in Table II.3. For an individual protein, the sum of all of the pair potentials of all of the interactions is computed, for example: $\Sigma PP = (PP[A][L] + PP[C][S] + 3PP[D][K] + 2PP[I][I] + ...)$ where there is one ALA-LEU interaction, one CYS-SER interaction, three ASP-LYS interactions, two ILE-ILE interactions, etc.. The individual spheres of the aromatic residues are treated separately. On average, a typical one-sphere residue made 2.8 contacts after two rounds of sphere growth, a two sphere aromatic residue made 5.1 contacts and tryptophan made 7.2 contacts.

3. Perturbation of Amino Acid Sequence and Side Chain Conformation

In order to test the extent of the dependence of side chain conformation on our packing algorithm, we used the program MIDAS (Ferrin *et al.*, 1988; Jarvis *et al.*, 1988) to substitute the crystallographically-determined side chain at each position of every protein in the data set with the same amino acid in its most frequently observed conformation as tabulated by Ponder and Richards (1987). We arbitrarily chose 180° for the outer dihedral angles (X_3 and greater) of methionine, arginine and lysine. Ponder and Richards did not compute rotamer conformations for the outer (X_3 +) dihedral angles of these residues because of small sample size (Met, Arg) and poor determination of surface residues (Lys, Arg). To create a control set of structures, we scrambled the sequence of each protein in the data set and, again using the most frequently occurring side chain conformations, built side chains onto the original

backbones. The program JUMBLE from the Baylor College Molecular Biology Information Resource (MBIR) suite of sequence analysis programs was used to scramble the sequences of the data set proteins.

4. Construction of Combinatorially-Folded Models

We subjected a set of 20 combinatorially-generated myoglobin models, eight flavodoxin models, and 10 adenylate kinase models to the packing evaluation. The myoglobin models were generated by identifying potential helix-helix packing sites on the A,B,E,F,G, and H helices and then creating an exhaustive set of "folded" structures incorporating inter-helix packing at these sites. Structures not satisfying chain connectivity constraints were discarded (Cohen *et al.*, 1979; Cohen & Sternberg, 1980b). Twenty structures remained at the end of this analysis. These structures fell into several different topological classes based on their helix packing geometry and root mean square (r.m.s.) deviation from the x-ray structure of sperm whale myoglobin (Takano, 1977). The flavodoxin and adenylate kinase structures were generated in a similar fashion, though the surviving structures did not fall into as many topological classes (Cohen *et al.*, 1982).

The combinatorial models do not include loop residues, so each myoglobin model has only 103, instead of the actual 153 residues, each flavodoxin has 93 instead of 138 residues and each adenylate kinase has 97 instead of 194 residues. The models contain coordinates for α-carbons only. We fitted ideal, all-backbone-atom α-helices or β-strands to the α-carbons in the models and then installed side chains in their most frequently observed conformations.

5. Construction of Random Three-Dimensional Structures

We generated several sets of random, compact three-dimensional structures in order to compare packing in these structures to packing in native proteins. These

structures were generated by employing a self-avoiding random walk. The walk is constrained to be within a sphere whose radius is equal to the radius a protein of a given molecular weight would have if it were exactly spherical. The constraining radius (R) is computed from the average density of globular proteins, 1.4 g/ml (Creighton, 1983). (See Cohen & Sternberg, 1980a.)

The correct amino acid sequence is used when building random structures and new amino acids are added sequentially, subject to certain constraints. Residue placement is recursive, so if a new residue *i* can not be placed after five attempts, residue *i-1* is repositioned. The structures contain only α-carbons and centroids β (and γ in the case of aromatic residues). Adjacent Cα's are placed 3.8 Å apart and all sets of three adjacent alpha carbons ($C\alpha_{i-1}$, $C\alpha_i$, $C\alpha_{i+1}$) are constrained to be at an angle of 106.3°. Centroids β lie in the plane of three adjacent Cα's at a distance that represents the average distance Cα to β for residues of that type in real structures. Aromatic *ring* centroids are placed at an angle of 110.5° to Cα and the Cβ. No two α-carbons or centroids are allowed to come closer than an absolute cutoff distance of 3.0Å from each other. Contact distances greater than or equal to 4.0 Å are allowed. If two "atoms" are within a distance of 3.0 to 4.0 Å, the position is either accepted or rejected using a Monte Carlo method. If the random number is larger than P, the new position is rejected. The resulting distribution of distances drops off as $1/r^3$, due to excluded volume effects. The absolute (3.0Å) and "soft" (4.0Å) cutoffs were chosen so as to resemble actual distributions of contact distances in the data set of structures.

Seven sets of 50 structures having the flavodoxin sequence were generated in this manner. Each set of structures had a different constraining radius. The most compact structures had a constraining radius of 100%R and the least compact had a radius of 160%R.

6. Energy Minimization of Flavodoxin

Flavodoxin (Smith *et al.*, 1977) was energy-minimized using AMBER (Singh *et al.*, 1987). We performed a total of 700 cycles of conjugate gradient minimization in intervals of 100 cycles. A distance-dependent dielectric constant ($\varepsilon = R$) and a united atom representation of the polypeptide chain were used (Weiner *et al.*, 1984). This calculation was conducted in the absence of flavine.

Results and Discussion

1. The Sphere Model for Amino Acid Packing

The representation of amino acids as spheres is a reasonable simplification of the packing problem. Spherical residue volume correlates well (correlation coefficient, r = 0.90) with amino acid volume as computed by Chothia (1975) using the Voronoi method. (See Figure 2.2.) Histidine ($V_{his} = 98$ Å^3) is an outlier for reasons that are unclear. It is possible that His is more compact than the other aromatic amino acids and thus actually better represented as a single sphere. The volume of tryptophan, when modeled as two spheres, is uncharacteristically small ($V_{trp} = 84$ Å^3) in comparison to its molecular weight and Voronoi volume. For this reason, we decided to represent tryptophan as three spheres: one at the β-carbon, one at the center of the five-membered ring and one at the center of the six-membered ring. When we recomputed the ideal radii for all residue types, incorporating this three-sphere model for tryptophan, the ideal radii of the remaining amino acids changed either not at all or by less than one percent. The volume of three sphere tryptophan is 110 Å^3.

The sphere growth calculation may be used both qualitatively and quantitatively to evaluate amino acid packing in model structures. Computer graphics provides a useful tool for visualizing the results of this calculation. Each amino acid in a protein can be colored according to the percentage of its ideal radius attained. This makes it possible

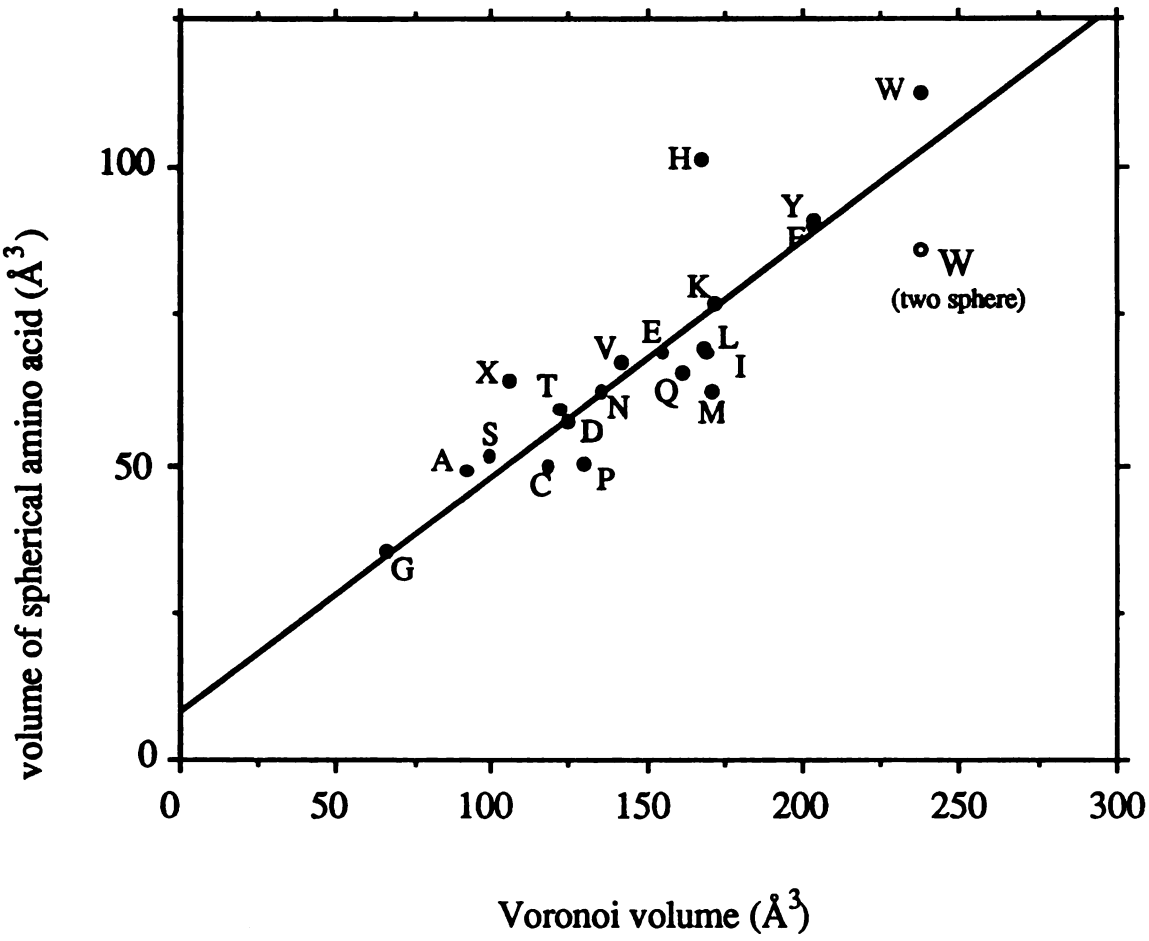


Figure 2.2 Volume of spherical amino acid (V_s) plotted as a function of amino acid volume as computed by C. Chothia using the method of Voronoi (V_s) (Chothia, 1974). Arginine is omitted since it is predominantly a surface residue and no accurate Voronoi volume for it could be computed. Two points are shown for tryptophan (W): one modeled as two spheres and the other as three spheres. The three sphere model is used for all calculations herein.
$V_s = 7.5 + 0.40V_v \quad r = 0.90$

to locate nonuniformities in packing. To display packing, we use the portion of the color wheel that includes the range green-cyan-blue-violet-magenta-red-orange. We chose this scheme because the extreme colors, green and orange, stand out well against the intermediate violets, making it easy to spot irregularities in packing. In

this scheme, cyan corresponds to a radius 60% ($r_i / r_i^0 = 0.6$) of the ideal size, red corresponds to 140% ($r_i / r_i^0 = 1.4$), and green represents 100% ($r_i = r_i^0$). A well-packed protein will be predominantly violet. A few exterior residues may be orange or red (loosely-packed) since they have fewer neighbors and are less likely to bump into other residues. Figure 2.3 shows four flavodoxin structures colored in this manner. The crystallographic structure is the most homogeneous while various models with increasing amounts of error appear inhomogeneous.

Homogeneity can be quantified through the statistical distribution of sphere sizes in an individual protein. For crystallographically-determined structures, this distribution is predominantly Gaussian but skewed slightly towards larger-than-ideal spheres: mean 100.0% of ideal; median 98.1%. The skewness corresponds to exterior residues (see Figure 2.4). The standard deviation of the distribution for an individual protein is typically ±13 percentage points (see Figure 2.5A.) There is no correlation between crystallographic resolution and mean sphere size nor crystallographic resolution and standard deviation of sphere sizes. Small proteins (< 100 residues) have the most deviant means and standard deviations.

A useful metric derived from this statistical characterization is the "number of outliers" (NO): the number of residues which achieve a final size greater than 112% or less than 88% of the ideal. The boundaries of 88% and 112% were determined so as to include 15% of residues in the data set on each side of the mean. For individual proteins, the number of outliers correlates well with molecular weight: $NO = -6.5 + 0.37N$; $r = 0.95$ (where N is the number of residues.)

The amino acid - amino acid pairwise interaction potentials derived from the number of centroid-centroid collisions are reasonable. Table II.4 ranks from best to worst which residues an amino acid of a particular type prefers to interact with according to the pair potential matrix (Table II.3). Arginine and lysine prefer to interact with glutamate and aspartate; serine and threonine prefer to interact with other hydrophilic

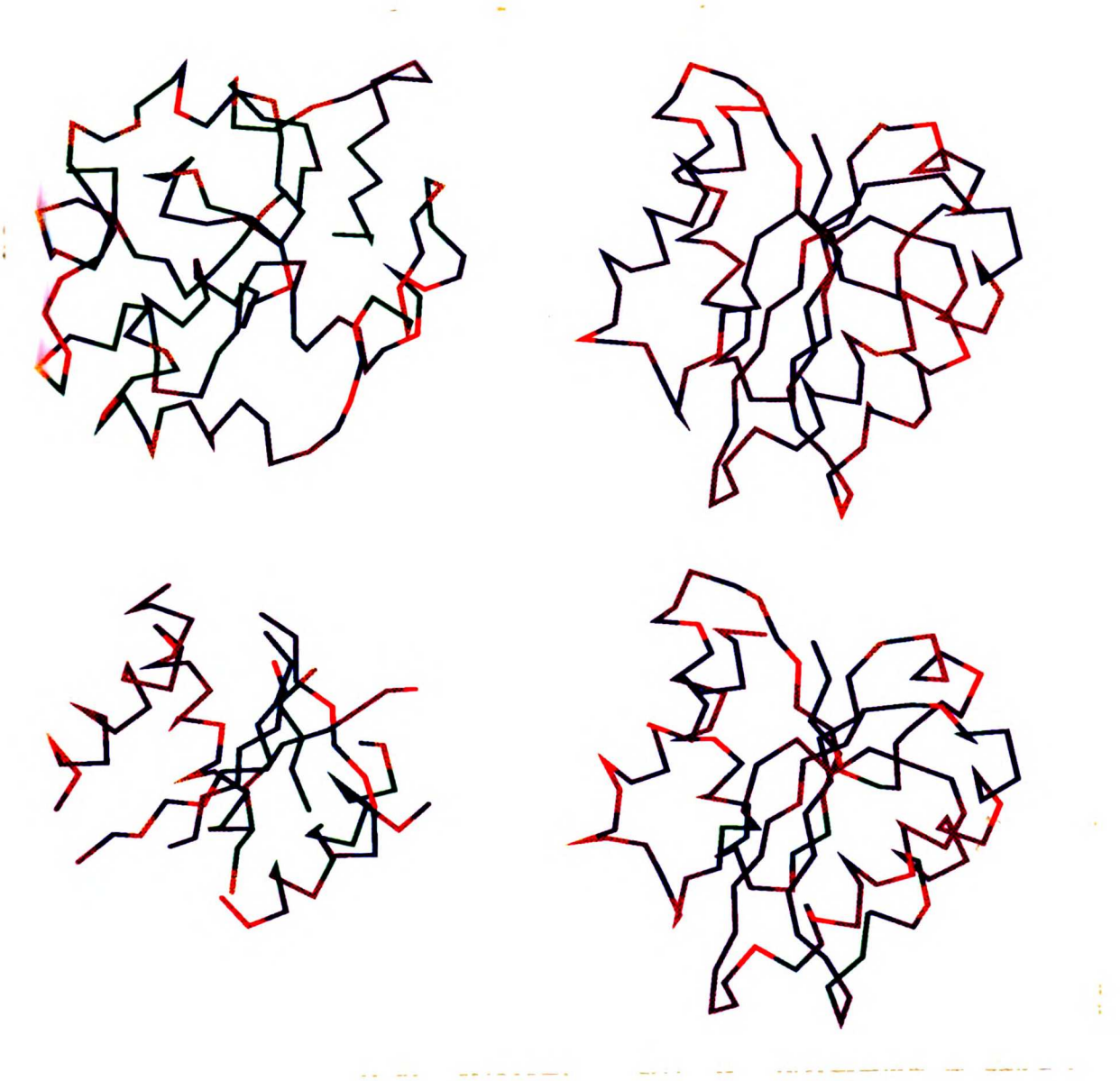


Figure 2.3 Graphical representation of packing for four models of flavodoxin. Only alpha carbons are shown. Residues are colored according to the percentage of their ideal sphere size they attain. Scale is green to orange where green residues are those which are too tightly packed ($r_i < r_i^0$) and orange residues are those which are too loosely packed ($r_i > r_i^0$). Purple is intermediate on this scale ($r_i \approx r_i^0$). Clockwise from top left: random walk flavodoxin; x-ray structure (Smith, *et al.*, 1977); x-ray structure with most common side chain dihedral angles installed; combinatorially-generated structure. The x-ray structure appears to be the most homogeneously colored and the random structure the least homogeneously colored. Both the combinatorially-generated model and the model with incorrect side chain dihedral angles contain packing errors. The combinatorially-produced model appears to have regional errors: the two helices on the left are too far apart while the two helices on the right collide with eachother and with the β-sheet. The model with incorrect side chain dihedral angles has packing errors distributed more evenly throughout the structure and is more uniformly well-packed than the combinatorially-generated structure.

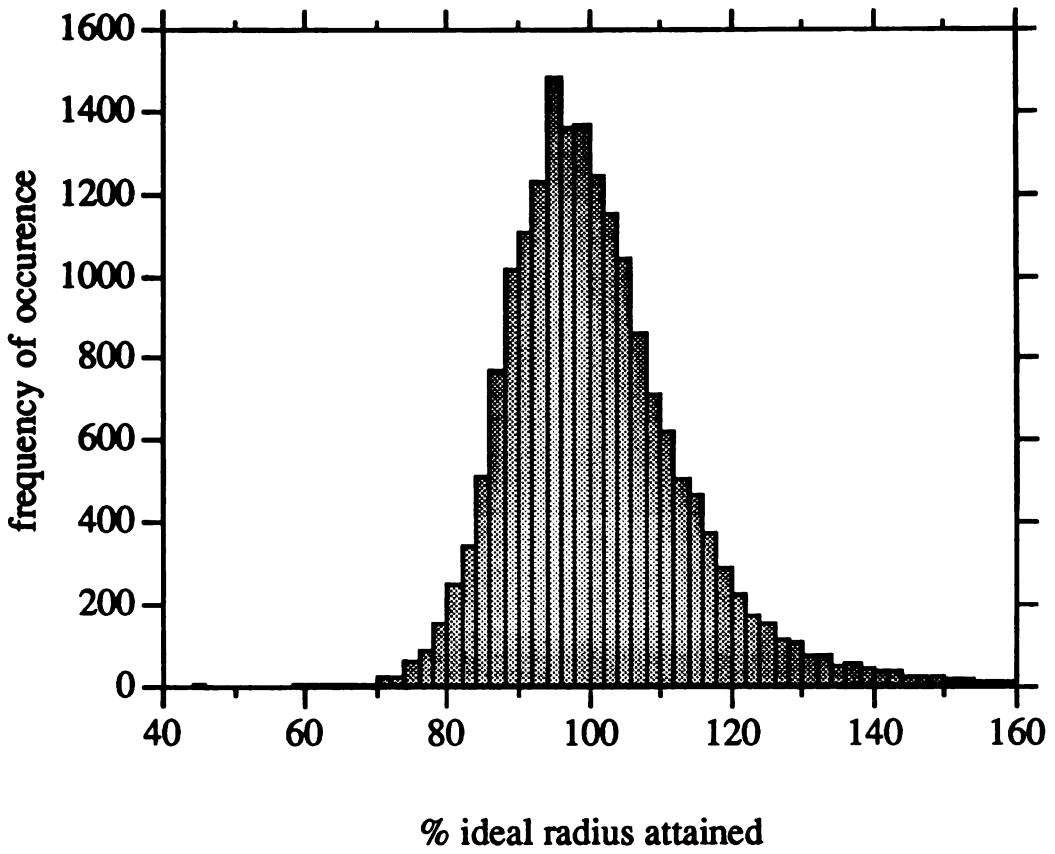


Figure 2.4. Distribution of sphere sizes relative to the ideal for amino acids in the data set of 72 proteins. Mean 100.0; standard deviation 13.4; median 98.1;

residues; the hydrophobic residues prefer to interact with other hydrophobic residues. A striking aberration is the tendency of cysteine and half-cystine to appear near the extremes of these rankings. This unusual behavior may be explained by the fact that the data set contains several iron binding proteins, such as 1FDX (ferrodoxin) and 4RXN (rubredoxin). These proteins bind iron by way of the sulfur atoms of cysteines. The presence of these proteins in the data set combined with the rarity of cysteine and half-cystine (Table II.1) may skew its statistics. Also, it is surprising that lysine and arginine interact favorably with the ring spheres of tryptophan. We looked at these interactions individually and concluded that they are legitimate. Tryptophan occasionally protrudes to the surface of proteins to make hydrogen bonds with surface

Table II.3

The pair potential matrix

	ALA	ARG	ASN	ASP	CYS	CYX	GLN	GLU	GLY	HIS	HIr	ILE	LEU	LYS	MET	PHE	PHr	PRO	SER	THR	TRP	TR5	TR6	TYR	TYr	VAL
ALA	-0.4	0.6	0.0	0.1	0.1	0.0	0.1	0.4	-0.1	0.0	0.7	-0.2	0.0	0.1	0.2	-0.2	0.2	0.2	-0.1	-0.1	0.0	0.2	0.3	0.2	0.2	-0.2
ARG	0.6	0.3	0.0	-0.7	0.5	-0.1	-0.3	-0.6	0.2	0.3	-0.2	0.0	0.1	0.8	0.1	0.0	-0.2	0.0	0.3	0.0	-0.3	-0.6	-0.7	0.1	-0.4	0.4
ASN	0.0	0.0	-0.3	-0.4	0.3	-0.2	-0.3	-0.2	-0.1	0.2	0.1	0.8	0.5	-0.2	0.8	0.3	0.4	-0.2	-0.3	0.1	0.1	0.4	0.7	-0.2	-0.2	0.3
ASP	0.1	-0.7	-0.4	0.0	0.5	0.1	0.1	0.0	-0.3	0.0	-0.1	0.5	0.6	-0.6	0.4	0.4	0.7	-0.1	-0.2	-0.3	0.9	0.7	1.3	0.3	0.3	0.5
CYS	0.1	0.5	0.3	0.5	-1.5	2.3	1.1	0.7	-0.4	-0.9	-0.7	-0.9	0.2	0.6	-0.5	-0.3	-0.2	-0.3	0.4	0.8	0.8	0.7	0.3	0.3	0.9	-0.1
CYX	0.0	-0.1	-0.2	0.1	2.3	-0.9	-0.2	0.5	-0.2	0.0	-0.3	0.8	0.1	0.9	0.9	0.6	0.2	0.2	-0.2	0.0	-0.1	-0.4	-0.8	-0.9	-0.4	0.3
GLN	0.1	-0.3	-0.3	0.1	1.1	-0.2	-0.1	-0.1	0.2	0.7	0.5	0.0	0.1	-0.3	0.1	0.5	-0.1	-0.3	-0.1	-0.2	0.4	0.2	0.1	0.3	-0.4	0.2
GLU	0.4	-0.6	-0.2	0.0	0.7	0.5	-0.1	-0.2	0.3	0.0	-0.5	0.3	0.5	-0.9	0.0	0.3	0.3	-0.1	-0.1	-0.2	0.7	0.3	0.7	0.3	0.1	0.5
GLY	-0.1	0.2	-0.1	-0.3	-0.4	-0.2	0.2	0.3	-0.2	0.0	0.8	0.1	0.2	0.2	0.3	0.0	0.9	0.1	-0.2	-0.2	-0.1	0.4	0.6	-0.2	0.2	-0.1
HIS	0.0	0.3	0.2	0.0	-0.9	0.0	0.7	0.0	0.0	-0.3	-0.5	0.3	0.2	0.1	0.9	-0.2	-0.2	-0.2	0.1	-0.1	0.5	-0.2	-0.7	-0.1	-0.4	0.1
HIr	0.7	-0.2	0.1	-0.1	-0.7	-0.3	0.5	-0.5	0.8	-0.5	-0.8	0.9	0.2	0.0	-0.1	0.0	-0.3	-0.3	-0.2	-0.2	0.1	-0.5	-0.5	0.3	-0.2	0.1
ILE	-0.2	0.0	0.8	0.5	-0.9	0.8	0.0	0.3	0.1	0.3	0.9	-0.4	-0.3	0.4	-0.4	-0.1	-0.2	0.1	0.5	0.2	-0.5	-0.5	-0.4	-0.1	0.1	-0.4
LEU	0.0	0.1	0.5	0.6	0.2	0.1	0.1	0.5	0.2	0.2	0.2	-0.3	-0.5	0.1	-0.3	-0.4	-0.4	0.1	0.3	0.5	-0.4	-0.4	-0.3	0.0	-0.1	-0.3
LYS	0.1	0.8	-0.2	-0.6	0.6	0.9	-0.3	-0.9	0.2	0.1	0.0	0.4	0.1	0.1	0.3	0.6	0.1	-0.1	0.1	0.1	0.4	0.1	-0.4	0.4	-0.4	0.4
MET	0.2	0.1	0.8	0.4	-0.5	0.9	0.1	0.0	0.3	0.9	-0.1	-0.4	-0.3	0.3	-1.0	0.3	-0.5	0.2	0.1	0.2	-0.3	-0.6	-0.7	0.0	0.1	-0.2
PHE	-0.2	0.0	0.3	0.4	-0.3	0.6	0.5	0.3	0.0	-0.2	0.0	-0.1	-0.4	0.6	0.3	-0.2	-0.8	0.2	0.3	0.2	-0.4	-0.4	-0.3	0.2	0.0	-0.1
PHr	0.2	-0.2	0.4	0.7	-0.2	0.2	-0.1	0.3	0.9	-0.2	-0.3	-0.2	-0.4	0.1	-0.5	-0.8	-0.6	0.1	0.6	0.8	-0.4	-0.6	-0.3	-0.3	0.1	-0.3
PRO	0.2	0.0	-0.2	-0.1	-0.3	0.2	-0.3	-0.1	0.1	-0.2	-0.3	0.1	0.1	-0.1	0.2	0.2	0.1	0.3	0.1	-0.1	0.0	0.3	-0.1	-0.2	-0.4	0.1
SER	-0.1	0.3	-0.3	-0.2	0.4	-0.2	-0.1	-0.1	-0.2	0.1	-0.2	0.5	0.3	0.1	0.1	0.3	0.6	0.1	-0.4	-0.3	0.5	0.6	0.6	0.1	0.5	0.3
THR	-0.1	0.0	0.1	-0.3	0.8	0.0	-0.2	-0.2	-0.2	-0.1	-0.2	0.2	0.5	0.1	0.2	0.2	0.8	-0.1	-0.3	-0.2	0.7	0.3	1.4	0.4	0.1	0.1
TRP	0.0	-0.3	0.1	0.9	0.8	-0.1	0.4	0.7	-0.1	0.5	0.1	-0.5	-0.4	0.4	-0.3	-0.4	-0.4	0.0	0.5	0.7	-0.9	0.3	-0.3	-0.5	0.3	-0.1
TR5	0.2	-0.6	0.4	0.7	0.7	-0.4	0.2	0.3	0.4	-0.2	-0.5	-0.5	-0.4	0.1	-0.6	-0.4	-0.6	0.3	0.6	0.3	0.3	2.3	-0.3	-0.4	0.8	-0.2
TR6	0.3	-0.7	0.7	1.3	0.3	-0.8	0.1	0.7	0.6	-0.7	-0.5	-0.4	-0.3	-0.4	-0.7	-0.3	-0.3	-0.1	0.6	1.4	-0.3	-0.3	-0.5	-0.4	0.4	0.0
TYR	0.2	0.1	-0.2	0.3	0.3	-0.9	0.3	0.3	-0.2	-0.1	0.3	-0.1	0.0	0.4	0.0	0.2	-0.3	-0.2	0.1	0.4	-0.5	-0.4	-0.4	-0.4	-0.2	0.1
TYr	0.2	-0.4	-0.2	0.3	0.9	-0.4	-0.4	0.1	0.2	-0.4	-0.2	0.1	-0.1	-0.4	0.1	0.0	0.1	-0.4	0.5	0.1	0.3	0.8	0.4	-0.2	-0.3	0.2
VAL	-0.2	0.4	0.3	0.5	-0.1	0.3	0.2	0.5	-0.1	0.1	0.1	-0.4	-0.3	0.4	-0.2	-0.1	-0.3	0.1	0.3	0.1	-0.1	-0.2	0.0	0.1	0.2	-0.3

Table II.4

The pair potential matrix sorted in order of preference of interaction (top → bottom = best → worst)

ALA ARG ASN ASP CYS CYX GLN GLU GLY HIS HIr ILE LEU LYS MET PHE PHr PRO SER THR TRP TR5 TR6 TYR TYr VAL

ALA ASP ASP ARG CYS CYX TYr LYS CYS CYS HIr CYS LEU GLU MET PHr PHE TYr SER ASP TRP ARG CYX CYX ARG ILE
ILE TR6 ASN LYS HIS TYR ARG ARG ASP TR6 CYS TRP PHE ASP TR6 LEU PHr CYS ASN SER ILE MET ARG TRP CYX LEU
PHE GLU GLN ASN ILE TR6 ASN HIr CYX HIr GLU TR5 PHr TR6 TR5 TRP TR5 GLN THR GLN TYR PHr HIS TR6 GLN PHr
VAL TR5 SER GLY HIr TR5 LYS ASN GLY TYr HIS ILE TRP TYr CYS TR5 MET HIr ASP GLU LEU HIr MET TR5 HIS VAL
GLY TYr CYX THR MET TYr PRO GLU SER HIS TR6 MET TR5 GLN PHr CYS LEU ASN CYX GLY PHE ILE HIr TYR LYS ALA
SER GLN GLU SER GLY HIr CYX THR THR PHE TR5 TR6 ILE ASN ILE TR6 TRP HIS GLY HIr PHr CYX TR6 PHr PRO MET
THR TRP LYS HIr PHE ASN THR GLN TYR PHr CYX VAL MET PRO LEU ALA HIr TYR HIr THR ARG LEU ILE ASN TYr TR5
ASN HIr PRO PRO PRO GLN GLN PRO ALA PRO PHr LEU TR6 HIr TRP HIS TR6 ASP ALA ALA MET PHE LYS GLY ASN CYS
CYX PHr TYR ASP PHr GLY GLU SER ASN TR5 PRO ALA VAL ALA VAL PHE TYR GLU GLN HIS TR6 TYR TYR PRO HIr GLY
HIS CYX TYr GLU VAL SER PHr ASP TRP THR ARG PHr TYr HIS HIr ILE VAL LYS GLU PRO CYX TR6 LEU TYr TYR PHE
LEU ASN GLY HIS ALA ARG SER HIS VAL TYR SER PHE ALA LEU GLU VAL ARG THR HIS ARG GLY HIS PHE HIS LEU TRP
TRP ILE ALA ALA LEU TRP ILE MET HIS ALA THR TYR TYR LYS TYR ARG CYS TR6 LYS CYX VAL VAL PHr ILE PHE TR6
ASP PHE ARG CYX ASN ALA ALA TYr PHE ASP TYr ARG ARG PHr ARG GLY HIS ARG MET ASN ALA LYS TRP LEU GLU HIS
CYS PRO HIr GLN TR6 HIS ASP GLY ILE CYX ASP GLN CYX SER GLN HIr ILE TRP PRO LYS PRO ALA TR5 MET ILE HIr
GLN THR THR TYR TYR THR LEU ILE PRO GLU MET GLY GLN THR SER TYr GLN GLY TYR TYr ASN GLN PRO ARG MET PRO
LYS LEU TRP TYr SER ASP MET PHE ARG GLY LYS PRO LYS TR5 TYr PRO LYS ILE ARG VAL HIr GLU VAL SER PHr THR
MET MET HIS MET ARG LEU TR6 PHr GLN LYS PHE TYr PRO GLY ALA THR PRO LEU LEU ILE TR5 PRO GLN VAL THR TYR
PHr TYR CYS PHE ASP PHr GLY TR5 LEU SER ASN THR CYS MET PRO TYR TYr PHr PHE MET TYr THR ALA ALA ALA GLN
PRO GLY PHE CYS LYS PRO TR5 TYR LYS VAL TRP GLU GLY ILE THR ASN ALA SER VAL PHE GLN TRP CYS PHE GLY TYr
TR5 ARG VAL ILE GLU VAL VAL ALA TYr ASN VAL HIS HIS TRP GLY GLU CYX VAL CYS TR5 LYS ASN TYr ASP VAL ASN
TYR HIS PHr VAL TR5 GLU TYR CYX GLU LEU LEU LYS HIr TYR LYS MET GLU ALA ILE TYR HIS GLY GLY CYS ASP CYX
TYr SER TR5 LEU THR PHE TRP LEU MET ARG TYR ASP SER VAL PHE SER ASN CYX TRP LEU SER SER SER GLN TRP SER
TR6 VAL LEU PHr TRP ILE HIr VAL TR5 ILE GLN SER ASN CYS ASP ASP SER MET TYr TRP GLU ASP ASN GLU TR6 ARG
GLU CYS TR6 TR5 TYr LYS PHE CYS TR6 TRP ALA ASN GLU PHE ASN GLN ASP PHE PHr CYS THR CYS GLU HIr SER LYS
ARG ALA ILE TRP GLN MET HIS TR6 HIr GLN GLY CYX THR ARG CYX CYX THR PRO TR6 PHr CYS TYr ASP LYS TR5 ASP
HIr LYS MET TR6 CYX CYS CYS TRP PHr MET ILE HIr ASP CYX HIS LYS GLY TR5 TR5 TR6 ASP TR5 THR THR CYS GLU

Footnote (for tables II.3 and II.4):
Nomenclature for the aromatic spheres: **HIS, PHE, TRP** and **TYR** correspond to the sphere located at the ß-carbon. **HIr, PHr, TYr** correspond to the sphere at the centroid of the aromatic ring. For tryptophan, **TR5** is the sphere at the centroid of the 5-membered ring and **TR6** is at centroid of the six-membered ring. CYX is half-cystine.

residues. Its rarity in the data set (237 tryptophans) may cause these interactions to be exaggerated. There are 20 instances of tryptophan and arginine being proximal and 33 instances of tryptophan and lysine being proximal. We chose to compute separate pair potentials for the individual spheres of the aromatic amino acids because the interactions they make are distinctly different. The centroid sphere of tyrosine, for example, prefers to interact with hydrophilic amino acids to a much greater extent than the β-carbon sphere. This is presumably due to the hydrogen bonding ability of tyrosine's hydroxyl group.

We decided to allow residues to contact both their first and second nearest neighbors in order to increase the "valence" of the amino acids. By restricting a residue to bump only into its nearest neighbor, information about the environment of that particular residue is limited and described by only one or perhaps two neighboring amino acids. In reality, an amino acid in a protein has several spatial neighbors. When two rounds of collision are performed, the situation is more realistic: a one-sphere residue makes contact with three neighboring spheres and an aromatic residue makes contact with between five and seven neighbors. When more interresidue interactions are allowed, the pair potential sum for an individual protein, particularly an unrefined model, is less likely to be dramatically influenced by a few chance interactions.

Our pair potential formulation is similar in its statistical nature to that of Miyazawa and Jernigan (1985), although we do not account for interaction with solvent. Tanaka and Scheraga (1976) derive pairwise interaction energies for all pairs of amino acids as well. Though the potentials derived by these three methods are qualitatively similar, our definition of an amino acid contact is unique since it is based on effective distances between residues rather than actual distances. We also include interactions between sequential nearest neighbors in the determination of a pair potential, whereas the other methods do not. As a consequence of these particularities, our matrix is best-suited to the analysis of structures presented here.

In addition to the identities of interacting residues, the sequential distance along the polypeptide chain between these residues is interesting. If a "short range" interaction is defined as one which occurs between residues which are five or fewer positions apart in sequence, 54% of the interactions occurring in the data set of 72 proteins are short-range.

3. Effect of Side Chain Conformation

Since the position of the centroid is computed from the positions of the side chain atoms, the dihedral angles (X_1-X_n) that the side chain adopts could affect the number of outliers and the pair potential sum. We investigated the dependence on side chain conformation by comparing three sets of structures: 1) the data set of 72 crystallographically-determined structures, 2) the same data set of 72 structures with the most frequently occurring side chain dihedral angles installed at every position, and 3) the data set of structures with the correct backbone but a scrambled amino acid sequence *and* side chains in their most frequently occurring conformations. The results of these calculations are shown in Figure 2.5.

Figures 2.5A shows mean sphere size, plotted as a function of standard deviation for the three sets of structures. The x-ray structures (circles) cluster about a mean (μ) of 100.4% of ideal radii and a standard deviation of (σ) 13.2%. The structures with incorrect side chain conformations (triangles) are less uniformly packed (μ = 97.4 %; σ = 21.2%) and the scrambled sequence structures are the least uniformly packed (μ = 97.1%; σ = 25.8%). The fact that the three classes of structures appear to fall into separate groups suggests that our method is sensitive to volume packing errors: the structures with merely incorrect side chain conformations should have mostly steric errors resulting from disrupted residue-residue interactions. The structures with a scrambled amino acid sequence, however, could have serious volume constraint violations where too many large residues occupy too small of a space.

In Figure 2.5B, NO is plotted as a function of N. Again, the crystallographically determined structures are more uniformly packed than either the scrambled sequence structures or the most frequent side chain conformation structures. The slopes of the lines indicate the percentage of residues expected to be outliers for a given set of structures. Thus, while the x-ray structures have 37% of their residues as outliers, the scrambled sequence typically have 62% and the most common side chain conformation structures have 55% of their residues as outliers. Even though the average standard deviations for the incorrect side chain conformation and scrambled sequence structures are quite different (Figure 2.5A), here they have similar numbers of outliers. This is because these two classes of structures are both too densely packed and have a large number of residues achieving a final size of less than 88% of their ideal size.
Figure 2.5C shows ΣPP plotted as a function of N. Here, the structures with merely incorrect side chain conformations are much more similar to the x-ray structures than are the scrambled sequence structures. The x-ray and incorrect side chain conformation structures usually attain a negative pair potential sum, in particular, only three x-ray structures have a pair potential sum greater than zero. The pairwise interactions that occur in the scrambled sequence structures are frequently destabilizing and result in a positive ΣPP. Here, only three structures have a pair potential sum *less* than zero. In Figure 2.5D, NO is plotted against ΣPP for the` three groups of structures. This graph suggests that taken together, the ΣPP and NO are discriminating measures of model quality since the three classes of structures fall into separate regions of the graph. The boundaries between the three classes of structures, however, are not easily demarcated. It may be possible to resolve the classes further using a third criterion of model quality.

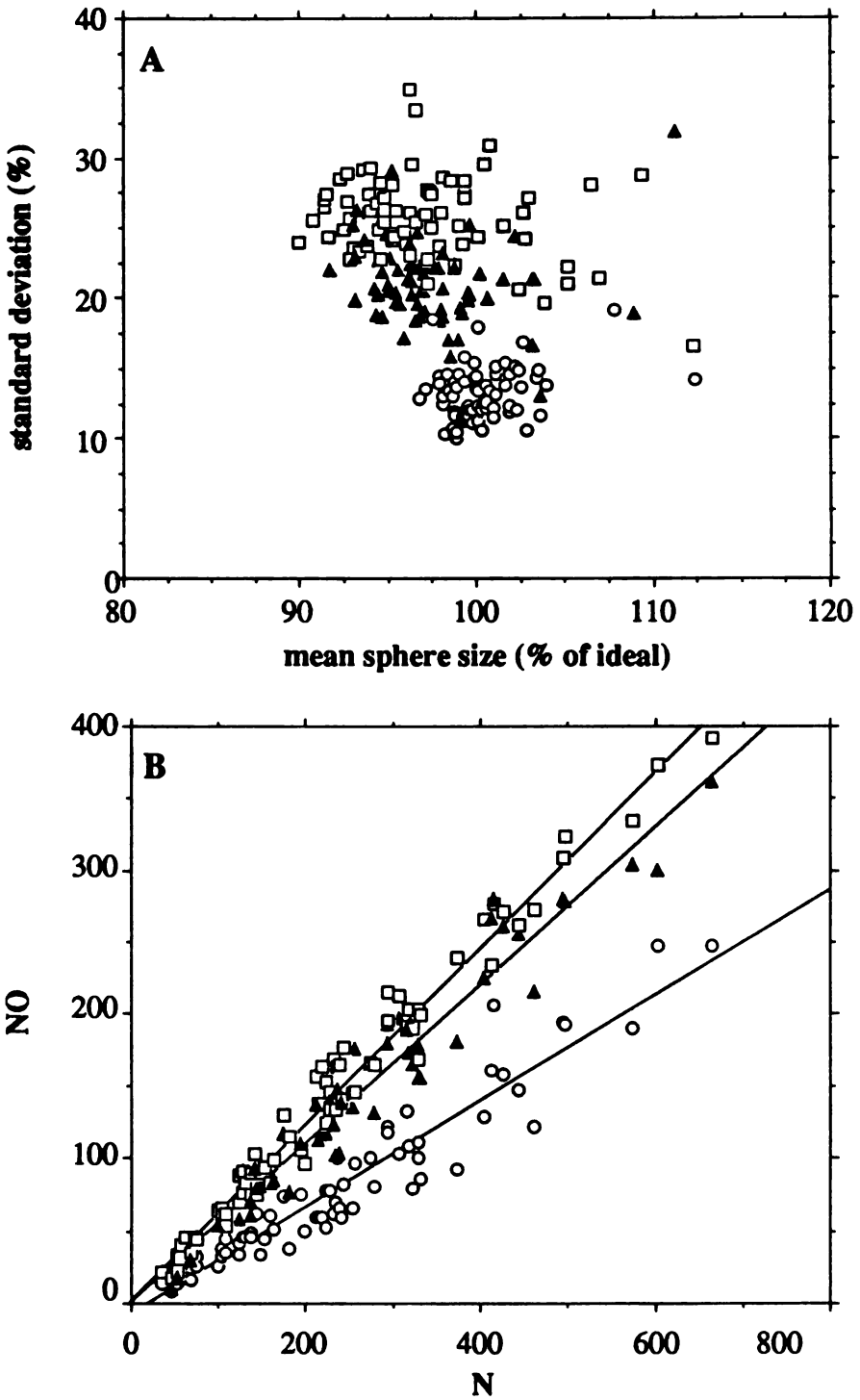


Figure 2.5 **A:** Packing uniformity of x-ray (open circle), scrambled sequence (open square) and perturbed side chain (filled triangle) structures: mean sphere size *versus* standard deviation of sphere sizes.
B: Number of outliers (NO) plotted against number of residues (N)

x-ray (circle):	NO = 0.37N - 6.50	r = 0.95
most frequently observed side chain (triangle):	NO = 0.55N + 0.12	r = 0.98
scrambled sequence (square):	NO = 0.62N + 0.68	r = 0.99

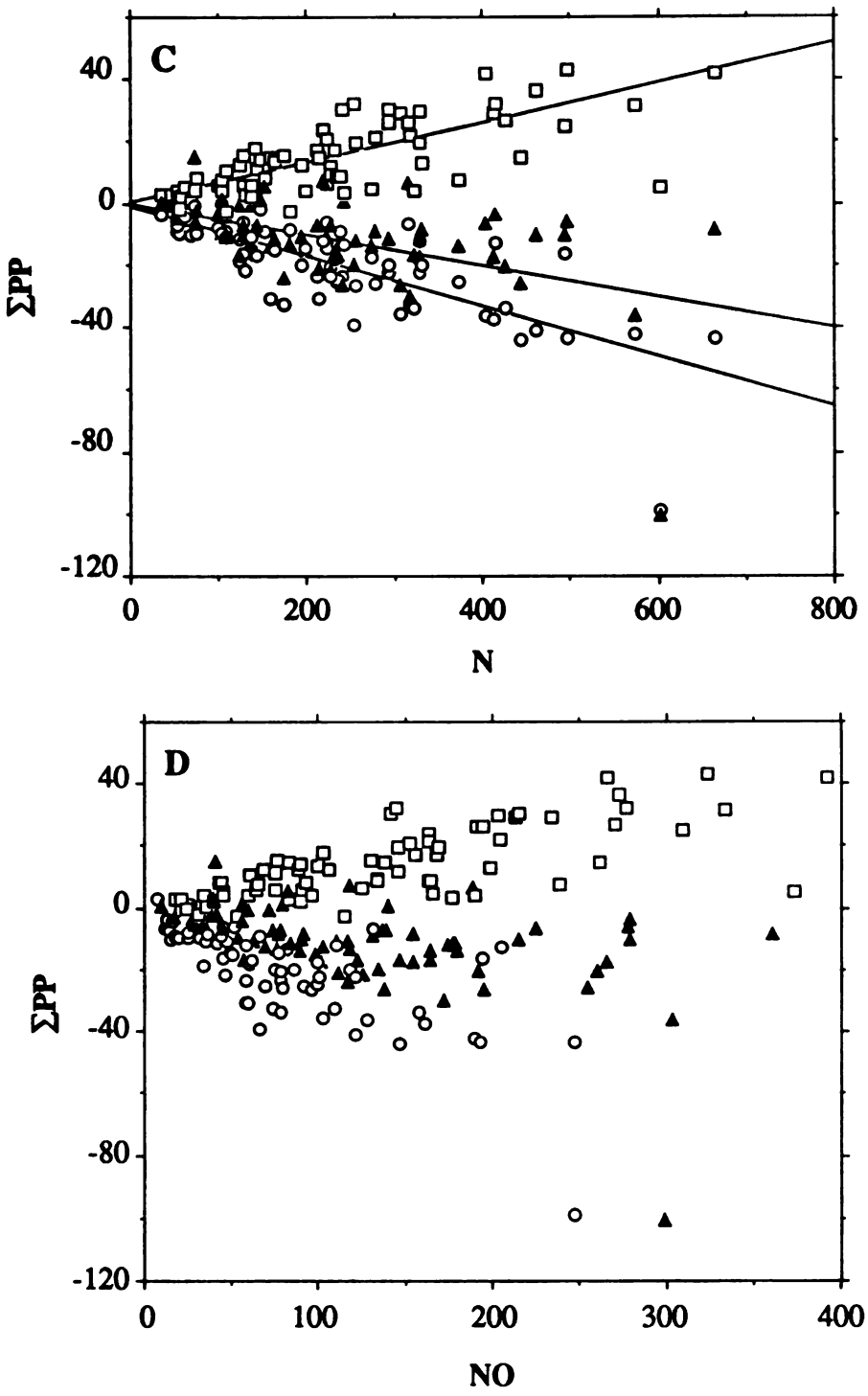


Figure 2.5 C: Pair potential sum (ΣPP) plotted against number of residues (N)
x-ray (circle): ΣPP = -0.080N - 1.3 r = -0.76
most frequently observed side chain (triangle): ΣPP = -0.051N + 0.2 r = -0.52
scrambled sequence (square): ΣPP = 0.057N + 1.2 r = 0.70
D: Pair potential sum (ΣPP) plotted against number of outliers (NO) (same symbols apply)

4. Combinatorially Folded Models

In the process of structure prediction, a modeler may generate a set of alternative model structures. To test the utility of our packing algorithm in discriminating between a group of proposed models, we subjected a set of combinatorially folded sperm whale myoglobin structures (Cohen, Richmond *et al.*, 1979) to our packing scheme. The results of this calculation are shown in Figures 2.6. The combinatorial models all have a large number of outliers (Figure 2.6A) compared to the native structure. However, the errors in model structures are comparable to a structure derived from the native backbone and statistically most likely side chain conformations. The models are a little too densely packed -- their mean sphere size is, on average, 97.5%. This is comparable to the data set structures with the most common side chain conformations installed. It is not possible to discriminate among the models on the basis of correctness of volume packing alone.

The pair potential sum suggests a way to differentiate between the models. In Figure 2.6B, ΣPP is plotted against r.m.s. deviation. There is a positive correlation ($r = 0.54$) between r.m.s. deviation and ΣPP. The most native-like structures have a negative pair potential sum and structure number thirteen, which has the lowest r.m.s. deviation, also has one of the lowest pair potential sums of all of models. Still, the pair potential sum is not uniformly instructive: structures nine and ten are relatively native-like but have pair potential sums near zero. Two poor models, numbers four and five, have pair potential sums even better than the best models. If these two structures are excluded from the calculation of a best fit line relating pair potential sum and r.m.s. deviation, the correlation improves to 0.75. It is not clear why these two structures have such favorable pair potential sums. The individual interactions that contribute to the pair potential sums of these two models are not particularly unusual, though both models have a fair number of interactions between oppositely-charged side chains. In general, these results suggest that sorting between alternative model structures may

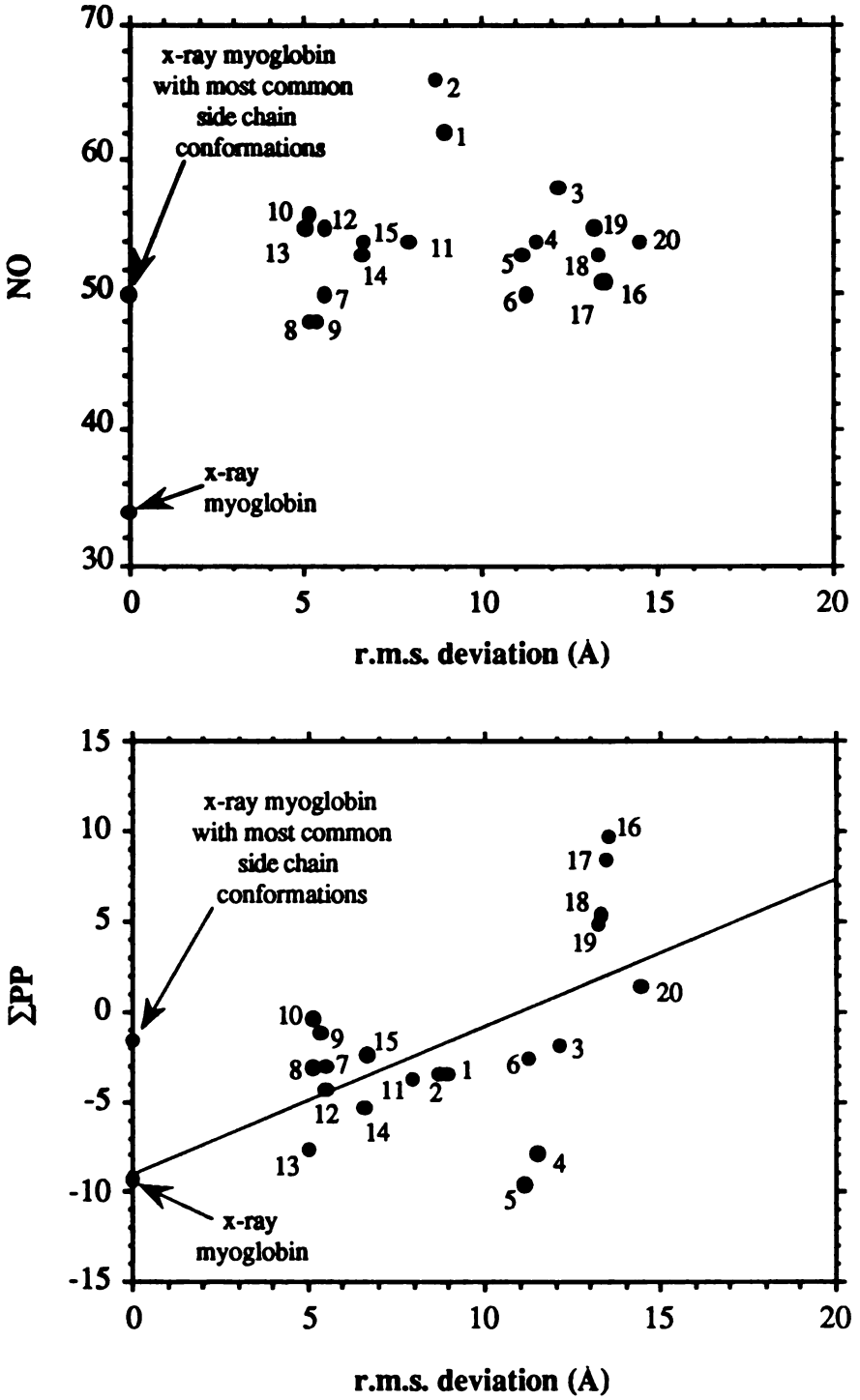


Figure 2.6 A: Number of outliers (NO) versus root-mean-square (r.m.s.) deviation for 20 combinatorially-folded myoglobins. **B:** Relationship between pair potential (ΣPP) and r.m.s. deviation for combinatorially-folded myoglobin models. R.m.s. deviation is computed using the coordinates of the alpha-carbons only. Best-fit line is computed using combinatorial structures only: $\Sigma PP = 0.81(r.m.s.) - 8.9$ $r = 0.54$ without two outliers (models 4 and 5): $\Sigma PP = 1.00(r.m.s.) - 9.7$ $r = 0.75$ The individual models are referenced 1-20 in each graph.

be facilitated by evaluating pair potential sums. A significant fraction of unreasonable structures can be discarded.

We also performed the packing calculation on a set of 10 combinatorially-generated adenylate kinase structures and eight flavodoxin structures. These proteins have similar α/β doubly-wound parallel sheet topologies (Richardson, 1981) Unlike the myoglobin models, there is much less variation in these structure sets. The range of r.m.s. deviations for the flavodoxins is 3.9Å to 4.5Å and the range for the adenylate kinases is 5.3Å to 6.7Å. The second best flavodoxin model (r.m.s. deviation = 3.95 Å) had the lowest pair potential sum and the worst model had the largest positive pair potential sum. However, a less optimal model (r.m.s. deviation = 4.35Å) also had a favorable pair potential sum. There was no correlation between r.m.s. deviation and ΣPP for the adenylate kinase models. There was little variation in the number of outliers in either set.

The pair potential sum does not do as well at discriminating between either the flavodoxin or adenylate kinase models as it does for the myoglobin models. The range of r.m.s. deviations for the flavodoxin and adenylate kinase models, however, is very small and hence it is difficult to discriminate between them using an approximate measure.

5. Effects of Energy Minimization on Packing

In the past, energy minimization techniques *in vacuo* have been observed to have a compressional effect on proteins because of the inadequacies of dielectric models and surface tension. We wanted to investigate whether these compressional effects could be detected using our measures of packing density and uniformity as reflected by the mean sphere size and standard deviation of sphere sizes. Table II.5 shows the outcome of successive steps of the minimization of flavodoxin.

As expected, with successive cycles of energy minimization, mean sphere size decreases, implying an overall compression of the molecule. Packing uniformity worsens -- both NO and standard deviation increase. Surprisingly, ΣPP worsens from -11.1 to a high of -4.8 after 200 cycles of minimization and then fluctuates about an

Table II.5

Congugate gradient minimization of flavodoxin

Cycles	RMS(Å)	E_{tot}	MEG	mean	SD	NO	ΣPP
0	0.00	1577.	144.9	102.5	13.6	46	-11.1
100	0.18	-2112.	122.9	102.8	13.8	48	-9.8
200	0.40	-2334.	179.1	101.8	14.3	52	-4.8
300	0.46	-2461.	42.6	101.4	14.0	54	-6.9
400	0.56	-2553.	30.9	101.0	14.3	51	-7.9
500	0.62	-2631.	64.9	100.6	14.4	62	-6.9
600	0.67	-2688.	26.3	100.2	14.5	60	-7.8
700	0.76	-2749.	93.5	99.4	14.6	70	-7.9

Footnote:

Cycles: number of cycles of conjugate gradient minimization
RMS: all-atom root-mean-square deviation from x-ray structure (Å)
E_{tot}: total energy (kcal/mol)
MEG: maximum energy gradient (kcal/(mol/Å))
mean: mean sphere size of residues in structure
SD: standard deviation of sphere sizes
NO: number of outliers
ΣPP: pair potential sum

average of -7.5. We did not notice any unusually large atomic displacements in the vicinity of the missing flavine ligand.

Whitlow and Teeter (1986) observe both compaction and molecular shape change over the course of minimization of crambin. In their study, molecular shape seemed to be most affected by the strength of electrostatic forces. The formation of new salt bridges and hydrogen bonds could occur in our models at the expense of atomic packing and lead to our observed decrease in packing uniformity.

These calculations have important implications for modeling. The modeler should be aware of the approximations inherent to minimizers and to use caution when refining a structure. Models built on the basis of sequence similarity to known structures may be the most susceptible to minimization artefacts. The inclusion of discrete solvent may avoid some of these artefacts (Richards *et al.*, 1989), as could the use of more sophisticated dielectric functions than $\varepsilon = r$ (Harvey, 1989).

6. Random Walk Structures

In order to understand and appreciate the difficulty in generating uniform packing, we created some "random" three-dimensional flavodoxin structures which were evaluated by QPACK. Six sets of random flavodoxins having different constraining radii, ranging from 100% to 160% of the expected protein radius (R), were generated using this method (see Table II.6.) It can be seen that as the constraining radius R is increased, the resulting random structures become less densely packed (mean radius increases.) Packing uniformity (standard deviation, NO) remains virtually constant. Not surprisingly, the pair potential sums of all but nineteen of 350 of these structures are positive (data not shown.)

The set of structures which behaves most like real proteins in terms of the average sequential distance between interacting residues is the 120% R set. In this set of structures, 54% of the interactions are between residues five or fewer apart in sequence -- the same as in the crystallographically-determined structure. The 120% R set of structures, however, are not densely packed (mean radius size = 103.7%.) The most densely packed set of structures is the 100% R set. Here, however, too many long-range interactions (> 5 residues apart in sequence) are forced and the percentage of short-range interactions is only 38%. In this set of structures, local interactions least resemble those observed in native proteins. These results point out that it is difficult to generate random three-dimensional structures which satisfy all characteristics of real proteins at once.

Table II.6

Statistics for random walk flavodoxins

R	RMS	mean	std.dev.	NO	%short range
100%	14.5 ± 1.1	98.3 ± 1.4	23.8 ± 1.2	99.0 ± 5.0	38 ± 3
110%	15.0 ± 1.3	101.5 ± 2.1	24.1 ± 1.5	99.1 ± 5.3	49 ± 5
120%	15.9 ± 1.2	103.7 ± 1.9	24.2 ± 1.8	100.3 ± 6.7	54 ± 5
130%	15.7 ± 1.0	105.0 ± 2.5	24.1 ± 1.4	99.4 ± 7.5	57 ± 6
140%	16.2 ± 1.4	106.2 ± 2.6	24.2 ± 1.5	100.4 ± 7.3	62 ± 6
150%	16.8 ± 1.5	109.2 ± 2.8	23.7 ± 1.5	101.6 ± 7.0	68 ± 6
160%	17.5 ± 1.4	109.0 ± 2.0	23.7 ± 1.6	101.5 ± 7.1	70 ± 6

Footnote:

Data for each R is averaged over a set of 50 random structures.
R: contstraining radius (100% = the radius flavodoxin would have if it were sphereical and of uniform density.)
RMS: root-mean-square deviation from x-ray structure
mean: mean sphere size
std.dev.: standard deviation of sphere sizes
NO: number of outliers
%short range: percent of reisdue collisions whichare between residues five or fewer positions apart in squence (computed over entire set of structures combined)

Conclusions

We have presented a new method for evaluating model protein structures. Amino acid packing, as evaluated by the sphere growth method, is consistent with the accuracy of model built structures: as one builds models which deviate more and more from crystallographically-determined structures, packing uniformity is seen to worsen, and amino acid distribution becomes less ideal. Structures with perturbed side chain conformations more closely resemble native proteins in terms of packing density and amino acid distribution than structures with scrambled amino acid sequences. Random walk structures do not resemble native proteins in many respects at all, aside from mean density.

Using the pair potential sum, which measures amino acid distribution, it is possible to eliminate a group of poor structures from a set of combinatorially-generated structures. The ability of the pair potential metric to discriminate between alternative combinatorially-generated structures is especially encouraging, since it could directly assist in structure prediction and perhaps lead to a second-generation structure prediction method which considers long-range interactions between amino acids distant in sequence, as well as local interactions. We are investigating the possibility that pair potential functions and their derivatives will provide information useful for model structure refinement.

Our packing algorithm may also be used as a graphical tool to assist in model-building. Since QPACK is fast, it should be feasible to incorporate it directly into a molecular modeling package. Once could envision constantly evaluating packing and coloring a model accordingly as modifications, such as side chain rotations or loop deformations, are made. We are planning to incorporate the QPACK algorithm in UCSF MIDAS (Ferrin, Huang *et al.*, 1988; Jarvis, Huang *et al.*, 1988) and its successors[†] .

The calculations involving structures with scrambled amino acid sequences and perturbed side chain conformations offer some insight into the number of sequences that can fit into a structural template. Given the large variety of protein sequences, yet the limited number of tertiary structural motifs (Richardson, 1981) it has been proposed that there exist broad "tertiary templates" for sequences which direct them to fold into a particular one of these motifs (Blundell & Sternberg, 1985; Ponder & Richards, 1987). Recent experiments by Lim and Sauer (1989) suggest that a number of sequences which are compatible with maintaining tertiary structure in a hydrophobic core of lambda repressor is governed primarily by amino acid composition, rather than combined amino acid volume. Steric clashes which result when side chains of an

[†] QPACK may now be accessed from Midas with the midas command "pdbrun".

appropriate total volume cannot efficiently pack around one another also play an important, but not easily experimentally quantifiable role. Our pair potential, which monitors amino acid composition, perhaps in conjunction with a rotamer searching algorithm (Ponder & Richards, 1987), to monitor steric interference, might be used as a preliminary screen to determine *a priori* which sequences can be accommodated in the core of a given protein.

Chapter 2 References

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. (1987). In *Crystallographic Databases -- Information Content, Software Systems, Scientific Applications*. (Allen, F.H., Bergeroff, G. & Sievers, R., ed.), pp. 107-132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Adman, E.T., Sieker, L.C. & Jensen, L.H. (1976). Structure Of Peptococcus Aerogenes Ferredoxin, Refinement At 2 Angstroms Resolution. *J. Biol. Chem.* **251**, 3801-3806.

Alden, R.A., Birktoft, J.J., Kraut, J., Robertus, J.D. & Wright, C.S. (1971). Atomic Coordinates For Subtilisin BPN (Or Novo). *Biochem. Biophys. Res. Comm.* . **45**, 337-344.

Arni, R., Heinemann, U., Maslowska, M., Tokuoka, R. & Saenger, W. (1988). Restrained Least-Squares Refinement Of The Crystal Structure Of The Ribonuclease T1*2'guanylic Acid Complex At 1.9 Angstroms Resolution. *J. Biol. Chem.* **263**, 15358-15368.

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.

Arutyunyan, E.G., Kuranova, I.P., Vainshtein, B.K. & Steigemann, W. (1980). X-ray Structural Investigation of Leghemoglobin. VI. Structure of Acetate-ferrileghemoglobin at a Resolution of 2.0 Angstroms. *Kristallografiya.* **25**, 80.

Babu, Y.S., Bugg, C.E. & Cook, W.J. (1988). Structure Of Calmodulin Refined At 2.2 Angstroms Resolution. *J. Mol. Biol.* **204**, 191-204.

Baker, E.N. (1988). Structure Of Azurin From Alcaligenes denitrificans. Refinement At 1.8 Angstroms Resolution And Comparison Of The Two Crystallographically Independent Molecules. *J. Mol. Biol.* **203**, 1071-1095.

Baker, E.N. & Dodson, E.J. (1980). Crystallographic Refinement Of The Structure Of Actinidin At 1.7 Angstroms Resolution By Fast Fourier Least-Squares Methods. *Acta Crystallogr., Sect.A.* **36**, 559-572.

Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C. & Wilson, I.A. (1976). Atomic Coordinates For Triose Phosphate Isomerase From Chicken Muscle. *Biochem. Biophys. Res. Com.* **72**, 146-155.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F.M., 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.

Betzel, C., Pal, G.P. & Saenger, W. (1988). Synchrotron X-ray Data Collection And Restrained Least-Squares Refinement Of The Crystal Structure Of Proteinase K At 1.5 Angstroms Resolution. *Acta Crystallogr., Sect.B.* **44**, 163-172.

Birktoft, J.J. & Banaszak, L.J. (1983). The Presence Of A Histidine-Aspartic Acid Pair In The Active Site Of 2-Hydroxyacid Dehydrogenases. X-ray Refinement Of Cytoplasmic Malate Dehydrogenase. *J. Biol. Chem.* **258**, 472-482.

Blake, C.C.F., Geisow, M.J., Oatley, S.J., Rerat, B. & Rerat, C. (1978). Structure Of Prealbumin, Secondary, Tertiary And Quaternary Interactions Determined By Fourier Refinement At 1.8 Angstroms. *J. Mol. Biol.* **121**, 339-356.

Blundell, T. & Sternberg, M.J.E. (1985). *Trends Biotechnol.* **3**, 228-235.

Blundell, T.L., Pitts, J.E., Tickle, I.J., Wood, S.P. & Wu, C. (1981). X-ray Analysis (1.4-Angstroms Resolution) Of Avian Pancreatic Polypeptide. Small Globular Protein Hormone. *Proc. Natl. Acad. Sci. USA.* **78**, 4175-4179.

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

Bode, W., Epp, O., Huber, R., Laskowski\0Jr., M. & Ardelt, W. (1985). The Crystal And Molecular Structure Of The Third Domain Of Silver Pheasant Ovomucoid (OMSVP3). *Eur. J. Biochem.* **147**, 387-395.

Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. & Kraut, J. (1982). Crystal Structures Of Escherichia coli And Lactobacillus casei Dihydrofolate Reductase Refined At 1.7 Angstroms Resolution. I. General Features And Binding Of Methotrexate. *J. Biol. Chem.* **257**, 13650-13662.

Borkakoti, N., Moss, D.S. & Palmer, R.A. (1982). Ribonuclease-A. Least-Squares Refinement Of The Structure At 1.45 Angstroms Resolution. *Acta Crystallogr., Sect.B.* **38**, 2210-2217.

Brunie, S., Bolin, J., Gewirth, D. & Sigler, P.B. (1985). The Refined Crystal Structure Of Dimeric Phospholipase A2 At 2.5 Angstroms. Access To A Shielded Catalytic Center. *J. Biol. Chem.* **260**, 9742-9749.

Bryant, T.N., Watson, H.C. & Wendell, P.L. (1974). The Structure Of Yeast Phospho-glycerate Kinase . *Nature.* **247**, 14-17.

Chothia, C. (1975). Structural Invariants in Protein Folding. *Nature (London).* **254**, 304-308.

Chou, P.Y. & Fasman, G.D. (1974). Prediction of Protein Conformation. *Biochemistry.* **13**, 222-244.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D. & Fletterick, R.J. (1983). Secondary Structure Assignment for α/β Proteins by a Combinatorial Approach. *Biochemistry*. **22**, 4894-4904.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D. & Fletterick, R.J. (1986a). Turn Prediction in Proteins Using a Pattern-Matching Approach. *Biochemistry*. **25**, 266-275.

Cohen, F.E., Kosen, P.A., Kuntz, I.D., Epstein, L.B., Ciardelli, T.L. & Smith, K.A. (1986b). Structure-Activity Studies of Interleukin-2. *Science*. **234**, 349-352.

Cohen, F.E., Richmond, T.J. & Richards, F.M. (1979). Protein Folding: Evaluation of some Simple Rules for the Assembly of Helices into Tertiary Structures with Myoglobin as an Example. *J. Mol. Biol.* **132**, 275-288.

Cohen, F.E. & Sternberg, M.J.E. (1980a). On the Prediction of Protein Structure: The Significance of the Root-mean-square Deviation. *J. Mol. Biol.* **138**, 321-333.

Cohen, F.E. & Sternberg, M.J.E. (1980b). On the Use of Chemically Derived Distance Constraints in the Prediction of Protein Structure with Myoglobin as an Example. *J. Mol. Biol.* **137**, 9-22.

Cohen, F.E., Sternberg, M.J.E. & Taylor, W.R. (1980). Analysis and Prediction of Protein β-Sheet Structures by a Combinatorial Approach. *Nature (London)*. **285**, 378-382.

Cohen, F.E., Sternberg, M.J.E. & Taylor, W.R. (1982). Analysis and Prediction of the Packing of α-Helices against a β-Sheet in the Tertiary Structure of Globular Proteins. *J. Mol. Biol.* **156**, 821-862.

Creighton, T.E. (1983). In *Proteins: Structures and Molecular Properties*. ed.), pp. 268, W. H. Freeman and Co., New York.

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

Dijkstra, B.W., Kalk, K.H. & Drenth, W.G.J.H.$J. (1981). Structure Of Bovine Pancreatic Phospholipase A2 At 1.7 Angstroms Resolution. *J. Mol. Biol.* **147**, 97-123.

Dreusicke, D., Karplus, P.A. & Schulz, G.E. (1988). Refined Structure Of Porcine Cytosolic Adenylate Kinase At 2.1 Angstroms Resolution. *J. Mol. Biol.* **199**, 359-371.

Eklund, H., Nordstrom, B., Zeppezauer, E., Soderlund, G., Ohlsson, I., Boiwe, T., Tapia, O. & Branden, C.-I. (1976). Three-dimensional Structure of horse Liver Alcohol Dehydrogenase at 2.4 Angstroms Resolution. *J. Mol. Biol.* **102**, 27-59.

Epp, O., Lattman, E.E., Schiffer, M., Huber, R. & Palm, W. (1975). The Molecular Structure Of A Dimer Composed Of The Variable Portions Of The Bence-jones Protein REI Refined At 2.0 Angstroms Resolution. *Biochemistry.* **14**, 4943-4952.

Fermi, G., Perutz, M.F., Shaanan, B. & Fourme, R. (1984). The Crystal Structure Of Human Deoxyhaemoglobin At 1.74 Angstroms Resolution. *J. Mol. Biol.* **175**, 159-174.

Ferrin, T.E., Huang, C.C., Jarvis, L.E. & Langridge, R. (1988). The MIDAS Display System. *J. Mol. Graphics.* **6**, 13-37.

Finzel, B.C., Poulos, T.L. & Kraut, J. (1984). Crystal Structure Of Yeast Cytochrome c Peroxidase Refined At 1.7-Angstroms Resolution . *J. Biol. Chem.* **259**, 13027.

Finzel, B.C., Weber, P.C., Hardman, K.D. & Salemme, F.R. (1985). Structure Of Ferricytochrome c(Prime) From Rhodospirillum molischianum At 1.67 Angstroms Resolution. *J. Mol. Biol.* **186**, 627-643.

Fita, I. & Rossmann, M.G. (1985). The NADPH Binding Site On Beef Liver Catalase. *Proc. Natl. Acad. Sci. USA.* **82**, 1604-1608.

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.

Fujinaga, M. & James, M.N.G. (1987). Rat Submaxillary Gland Serine Protease, Tonin. Structure Solution And Refinement At 1.8 Angstroms Resolution. *J. Mol. Biol.* **195**, 373-396.

Garnier, J., Osguthorpe, D.J. & Robson, B. (1978). *J. Mol. Biol.* **120**, 97-120.

Greer, J. (1981). Comparative Model-Building of the Mammalian Serine Proteases. *J. Mol. Biol.* **153**, 1027-1042.

Greer, J. (1985). Model Structure for the Inflammatory Protein C5a. *Science.* **228**, 1055-1060.

Guss, J.M. & Freeman, H.C. (1983). Structure Of Oxidized Poplar Plastocyanin At 1.6 Angstroms Resolution. *J. Mol. Biol.* **169**, 521-563.

Harvey, S.C. (1989). Treatment of Electrostatic Effects in Macromolecular Modeling. *Proteins: Struct. Func. Gen.* **5**, 78-92.

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.

Holmes, M.A. & Matthews, B.W. (1982). Structure Of Thermolysin Refined At 1.6 Angstroms Resolution. *J. Mol. Biol.* **160**, 623-639.

James, M.N.G. & Sielecki, A.R. (1983). Structure And Refinement Of Penicillopepsin At 1.8 Angstroms Resolution. *J. Mol. Biol.* **163**, 299-361.

Jarvis, L., Huang, C., Ferrin, T. & Langridge, R. (1988). UCSF MIDAS: Molecular Interactive Display And Simulation. *J. Mol. Graphics.* **6**, 2-27.

Kamphuis, I.G., Kalk, K.H., Swarte, M.B.A. & Drenth, J. (1984). Structure Of Papain Refined At 1.65 Angstroms Resolution. *J. Mol. Biol.* **179**, 233-256.

Karplus, P.A. & Schulz, G.E. (1987). Refined Structure Of Glutathione Reductase At 1.54 Angstroms Resolution. *J. Mol. Biol.* **195**, 701-729.

Legg, M.J. (1977). Thesis, Texas Agricultural and Mechanical University.

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.

Levitt, M. (1976). A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding. *J. Mol. Biol.* **104**, 56-107.

Levitt, M. & Warshel, A. (1975). Computer Simulation of Protein Folding. *Nature*. **254**, 694-698.

Lim, V.I. (1974a). Algorithms for Prediction of α-Helical and β-Structural Regions in Globular Proteins. *J. Mol. Biol.* **88**, 873-894.

Lim, V.I. (1974b). Structural Principles of the Globular Organization of Protein Chains. A Stereochemical Theory of Globular Protein Secondary Structure. *J. Mol. Biol.* **88**, 857-872.

Lim, W.A. & Sauer, R.T. (1989). Alternative Packing Arrangements in the Hydrophobic Core of Lambda Repressor. *Nature*. **339**, 31-36.

Love, R.A. & Stroud, R.M. (1986). The Crystal Structure Of Alpha-bungarotoxin At 2.5 Angstroms Resolution. Relation To Solution Structure And Binding To Acetylcholine Receptor. *Protein Eng*. **1**, 37.

Marquart, M., Deisenhofer, J., Huber, R. & Palm, W. (1980). Crystallographic Refinement And Atomic Models Of The Intact Immunoglobulin Molecule Kol And Its Antigen-binding Fragment At 3.0 Angstroms And 1.9 Angstroms Resolution. *J.Mol.Biol.* **141**, 369-391.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). The Geometry Of The Reactive Site And Of The Peptide Groups In Trypsin, Trypsinogen And Its Complexes With Inhibitors. *Acta Crystallogr., Sect.B*. **39**, 480-490.

Meyer, E., Cole, G., Radahakrishnan, R. & Epp, O. (1988). Structure Of Native Porcine Pancreatic Elastase At 1.65 Angstroms Resolution. *Acta Crystallogr., Sect.B* . **44**, 26-38.

Miyazawa, S. & Jernigan, R.L. (1985). Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules* . **18**, 534-552.

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**, 201-225.

Moult, J., Sussman, F. & James, M.N.G. (1985). Electron Density Calculations As An Extension Of Protein Structure Refinement. Streptomyces griseus Protease At 1.5 Angstroms Resolution. *J.Mol.Biol.* **182**, 555-566.

Novotny, J., Bruccoleri, R. & Karplus, M. (1984). An Analysis of Incorrectly Folded Protein Models. *J. Mol. Biol.* **177**, 787-818.

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

Pletnev, V.Z., Kuzin, A.P. & Malinina, L.V. (1982). Actinoxanthin Structure at the Atomic Level. *Bioorg. Khim.* **8**, 1637.

Ploegman, J.H., Drent, G., Kalk, K.H. & Hol, W.G.J. (1978). Structure Of Bovine Liver Rhodanese. I. Structure Determination At 2.5 Angstroms Resolution And A Comparison Of The Conformation And Sequence Of Its Two Domains. *J. Mol. Biol.* **123**, 557-594.

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.

Poulos, T.L., Finzel, B.C. & Howard, A.J. (1987). High-resolution Crystal Structure Of Cytochrome P450cam. *J. Mol. Biol.* **195**, 687-700.

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.

Reeke, G.N., Becker, J.W. & Edelman, G.M. (1975). The Covalent And Three-dimensional Structure Of Concanavalin A, iv.Atomic Coordinates,Hydrogen Bonding,And Quaternary Structure. *J. Biol. Chem.* **250**, 1525-1547.

Rees, D.C., Lewis, M. & Lipscomb, W.N. (1983). Refined Crystal Structure Of Carboxypeptidase A At 1.54 Angstroms Resolution. *J. Mol. Biol.* **168**, 367-387.

Remington, S.J., Woodbury, R.G., Reynolds, R.A., Matthews, B.W. & Neurath, H. (1988). The Structure Of Rat Mast Cell Protease II At 1.9-Angstroms Resolution. *Biochemistry*. **27**, 8097-8105.

Richards, F.M. (1974). The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density. *J. Mol. Biol.* **82**, 1-14.

Richards, F.M. (1977). Areas, Volumes, Packing and Protein Structure. *Ann. Rev. Biophys. Bioeng.* **6**, 151-176.

Richards, W.G., King, P.M. & Reynolds, C.A. (1989). Solvation Effects. *Protein Eng.* **2**, 319-327.

Richardson, J.S. (1981). The Anatomy and Taxonomy of Protein Structure. *Adv. Protein Chem.* **34**, 167-339.

Satyshur, K.A., Rao, S.T., Pyzalska, D., Drendel, W., Greaser, M. & Sundaralingam, M. (1988). Refined Structure Of Chicken Skeletal Muscle Troponin C In The Two Calcium State At 2 Angstroms Resolution. *J. Biol. Chem.* **263**, 1628-1647.

Saul, F.A., Amzel, L.M. & Poljak, R.J. (1978). Preliminary refinement and structural analysis of the Fab fragment from Immunoglobulin New at 2.0 Å resolution. *J. Biol. Chem.* **253**, 585.

Singh, U.C., Weiner, P.K., Caldwell, J. & Kollman, P.A. (1987). AMBER 3.0. University of California, San Fransicso.

Smith, J.L., Corfield, P.W.R., Hendrickson, W.A. & Low, B.W. (1988). Refinement At 1.4 Angstroms Resolution Of A Model Of Erabutoxin b. Treatment Of Ordered Solvent And Discrete Disorder. *Acta Crystallogr., Sect.A.* **44**, 357-368.

Smith, W.W., Burnett, R.M., Darling, G.D. & Ludwig, M.L. (1977). Structure Of The Semiquinone Form Of Flavodoxin From Clostridium Mp. Extension Of 1.8 Angstroms Resolution And Some Comparisons With The Oxidized State. *J. Mol. Biol.* **117**, 195-225.

Steigemann, W. & Weber, E. (1979). Structure Of Erythrocruorin In Different Ligand States Refined At 1.4 Angstroms Resolution. *J. Mol. Biol.* **127**, 309-338.

Strynadka, N.C.J. & James, M.N.G. (1988). Two Trifluoperazine binding Sites on Calmodulin Predicted from Comparative Molecular Modeling withe Troponin-C. *Proteins: Structure, Function and Genetics.* **3**, 1-17.

Suguna, K., Padlan, E.A., Smith, C.W., Carlson, W.D. & Davies, D.R. (1987). Binding Of A Reduced Peptide Inhibitor To The Aspartic Proteinase From Rhizopus chinensis. Implications For A Mechanism Of Action. *Proc. Natl. Acad. Sci. USA.* **84**, 7009-7013.

Summers, N.L., Carlson, W.D. & Karplus, M. (1987). Analysis of Side-chain Orientations in Homologous Proteins. *J. Mol. Biol.* **196**, 175-198.

Szebenyi, D.M.E. & Moffat, K. (1986). The Refined Structure Of Vitamin D-dependent Calcium-binding Protein From Bovine Intestine. Molecular Details, Ion Binding, And Implications For The Structure Of Other Calcium-binding Proteins. *J. Biol. Chem.* **261**, 8761-8777.

Tainer, J.A., Getzoff, E.D., Beem, K.M., Richardson, J.S. & Richardson, D.C. (1982). Determination And Analysis Of The 2 Angstrom Structure Of Copper, Zinc Superoxide Dismutase. *J. Mol. Biol.* **160**, 181-217.

Takano, T. (1977). *J. Mol. Biol.* **110**, 569-584.

Takano, T. & Dickerson, R.E. (1980). Redox Conformation Changes In Refined Tuna Cytochrome c. *Proc. Natl. Acad. Sci. USA.* **77**, 6371-6375.

Tanaka, S. & Scheraga, H.A. (1976). Medium- and Long-Range Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. *Macromolecules* . **9**, 945-950.

Taylor, W.R. (1986). Identification of Protein Sequence Homology by Consensus Template Alignment. *J. Mol. Biol.* **188**, 233-258.

Taylor, W.R. (1988). Review of Pattern Matching Methods in Protein Sequence Comparison and structure Prediction. *Protein Eng.* **2**, 77-86.

Terwilliger, T.C. & Eisenberg, D. (1982). The Structure Of Melittin. I. Structure Determination And Partial Refinement. *J. Biol. Chem.* **257**, 6010-6015.

Tsukada, H. & Blow, D.M. (1985). Structure Of Alpha-Chymotrypsin Refined At 1.68 Angstroms Resolution. *J. Mol. Biol.* **184**, 703-711.

Vijay-Kumar, S., Bugg, C.E. & Cook, W.J. (1987). Structure Of Ubiquitin Refined At 1.8 Angstroms Resolution. *J. Mol. Biol.* **194**, 531-544.

Watenpaugh, K.D., Sieker, L.C. & Jensen, L.H. (1980). Crystallographic Refinement Of Rubredoxin At 1.2 Angstroms Resolution. *J. Mol. Biol.* **138**, 615-633.

Weaver, L.H. & Matthews, B.W. (1987). Structure Of Bacteriophage T4 Lysozyme Refined At 1.7 Angstroms Resolution. *J. Mol. Biol.* **193**, 189-199.

Weber, I.T. & Steitz, T.A. (1987). Structure Of A Complex Of Catabolite Gene Activator Protein And Cyclic AMP Refined At 2.5 Angstroms Resolution. *J. Mol. Biol.* **198**, 311-326.

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.

White, J.L., Hackert, M.L., Buehner, M., Adams, M.J., Ford, G.C., Jr., P.J.L., Smiley, I.E., Steindel, S.J. & Rossmann, M.G. (1976). A Comparison Of The Structures Of Apo Dogfish M4 Lactate Dehydrogenase And Its Ternary Complexes. *J. Mol. Biol.* **102**, 759-779.

Whitlow, M. & Teeter, M.M. (1986). An Empirical Examination of Potential Energy Minimization Using the Well-Defined Structure pf the Protein Crambin. *J. Am. Chem. Soc.* **108**, 7163-7172.

Zvelebil, M.J., Barton, G.J., Taylor, W.R. & Sternberg, M.J.E. (1987). Prediction of Protein Secondary Structure and Active Sites Using Alignment of Homologous Sequences. *J. Mol. Biol.* **195**, 952-961.

Chapter 3.
Protein Folding: Effect of Packing Density on Chain Conformation†

† This chapter is in press in the Journal of Molecular Biology and should be published in 1991.

Introduction

Fifty to sixty percent of amino acids in a globular protein participate in some form of secondary structure, typically as α-helices or β-sheets. While several forces are predicted to stabilize such structures, the recent work of Chan and Dill (1989; 1990b) suggests that compactness may be a significant driving force for the formation of secondary structure in globular proteins. The origin of the compacting force is postulated to be the hydrophobic effect which seeks to minimize exposed non-polar surface area in water. These workers have enumerated the conformational states available to a chain on a cubic lattice and shown that in compact conformations, about 50% of the residues participate in some form of secondary structure.

A major advantage of lattice models is computational efficiency -- simulations on lattices involve integer arithmetic rather than computationally costly floating point operations. Excluded volume effects are handled seamlessly. In addition, exhaustive searches of conformation space on a lattice are feasible. Several types of lattices have been used to study protein folding. Square and cubic lattices have been used by Go and co-workers to simulate protein folding using a Monte Carlo method (Go & Taketomi, 1978). Covell and Jernigan and have investigated face-centered cubic lattices and body-centered cubic lattices for representing proteins (Covell & Jernigan, 1990) and Skolnick and co-workers have used both a tetrahedral lattice (Skolnick & Kolinski, 1989) and a 2-1-0, or "knight's walk" lattice in their dynamics simulations (Skolnick & Kolinski, 1990). Each of these lattices has advantages over the others in terms of how realistically it reproduces some aspect of protein structure, or in how efficiently it allows one to explore conformational space. While lattices are a reasonable approach to studying the properties of protein molecules, a critical assessment of their advantages and limitations has not been undertaken. In this paper we address the effects of lattice constraints, compactness constraints, and shape constraints on chain fold and discuss their relevance to understanding protein folding.

We have tested the Chan-Dill hypothesis for proteins using a model of the polypeptide chain which is not restricted to a lattice. Structures are generated via a self-avoiding walk inside a constraining ellipsoid. Virtual bond angles and dihedral angles between sequential alpha-carbons are chosen at random from the distribution of angles observed in native proteins. We call these "native-like random walk structures." We are able to make large numbers of structures which obey the same boundary constraints and have the same distribution of virtual bond angles and dihedral angles between α-carbons seen in protein structures. The shape and density of a set of structures are easily varied. However, this representation does not allow us to enumerate all possible structures, so complete sampling cannot be guaranteed.

We find qualitative agreement with the conclusion of Chan and Dill that compactness induces secondary structure formation. However, a significant discrepancy exists as to the magnitude of this effect. In folding a protein from an unconstrained random walk to the density of typical proteins, secondary structure does not increase significantly (< 5%). To obtain a significant increase in secondary structure content, the protein must be packed more than 20% more densely than most protein structures are folded. The secondary structure that does form is predominantly alpha-helical.

Shape can also influence secondary structure content. Helical structure is enhanced in random walk structures constrained to extremely prolate ellipsoids. Extremely oblate ellipsoids contain significantly less helical structure, but contrary to our expectations, little strand-strand pairing is found. The lack of sheet structure in our simulations suggests that the cubic lattice imposes significant biases on the types of structures that are observed.

Methods

1. Generation of random walk structures

The native-like random walk structures are constructed by performing self-avoiding random walks. Consecutive alpha-carbons are separated by 3.8 Å and joined by virtual bonds. This mimics the distance between sequential alpha-carbons in native proteins. The model construction method is shown in Figure 3.1. Each structure is generated starting from a random position within the constraining ellipsoid. After three atoms have been placed, the position of each successive atom is determined by randomly choosing an inter-Cα angle α ($\angle C\alpha_{i-1}\ C\alpha_i\ C\alpha_{i+1}$) and a torsion angle τ ($\angle C\alpha_{i-2}\ C\alpha_{i-1}\ C\alpha_i\ C\alpha_{i+1}$) subject to the correlated α/τ probability distribution observed in native proteins (Figure 3.2). The frequency with which a pair of angles (α, τ) is picked is directly related to the frequency with which that combination occurs in native proteins. The α/τ distribution has a direct mapping to ϕ/ψ space normally used to describe backbone geometry in proteins and follows from the local excluded volume effects of a polypeptide chain. The distribution of angles ϕ and ψ derived from crystal structures is very similar to the one predicted from a hard-sphere model by Ramachandran et al. (1963; 1968) Unlike torsion angles ϕ and ψ, however, angles α and τ have equivalents in lattice models. A similar distribution was derived previously by Levitt (1975). For computational efficiency in choosing (α,τ) pairs, the distribution was digitized into 10368 2.5° x 2.5° bins and then mapped to one dimension and integrated as described in Numerical Recipes in C (Press et al., 1988, page 215). Therefore, angles smaller than 75°, which presumably are the result of experimental error, are never chosen.

The self-avoidance procedure works in the following way. A new alpha-carbon *i* is always allowed to be placed at least some minimum distance d_{min} (4.25 Å) from any other non-bonded alpha-carbon *j* ($j > i + 2$) but never closer than an absolute cutoff distance, d_{abs} (3.75 Å). If an atom position is chosen that places it between d_{abs} and

d_{min} of another atom, a Monte Carlo approach is taken to decide whether to allow placement. A random number between 0 and 1 is chosen. If that number is smaller than μ where

$$(1) \quad \mu = 1 - \frac{d_{min} - d_{ij}}{d_{min} - d_{abs}} \qquad \text{(where } d_{ij} \text{ is the distance between atoms i and j)}$$

then placement is allowed. Thus, distances closer to d_{min} are usually accepted and distances closer to d_{abs} are usually rejected. This soft sphere model was found to be more realistic than a hard sphere model. The distances of 3.75 Å and 4.25 Å where chosen to most closely approximate the distribution of inter-Cα distances in native proteins. Atom placement is recursive. If an atom cannot be placed after four

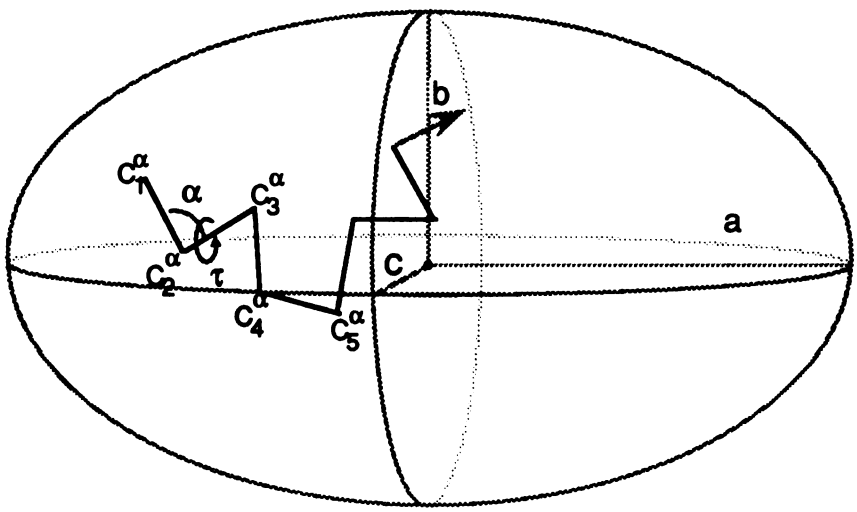


Figure 3.1. Construction of native-like random walk proteins showing constraining ellipsoid with vertices a, b, and c. Virtual bond angle α and virtual dihedral angle τ are indicated.

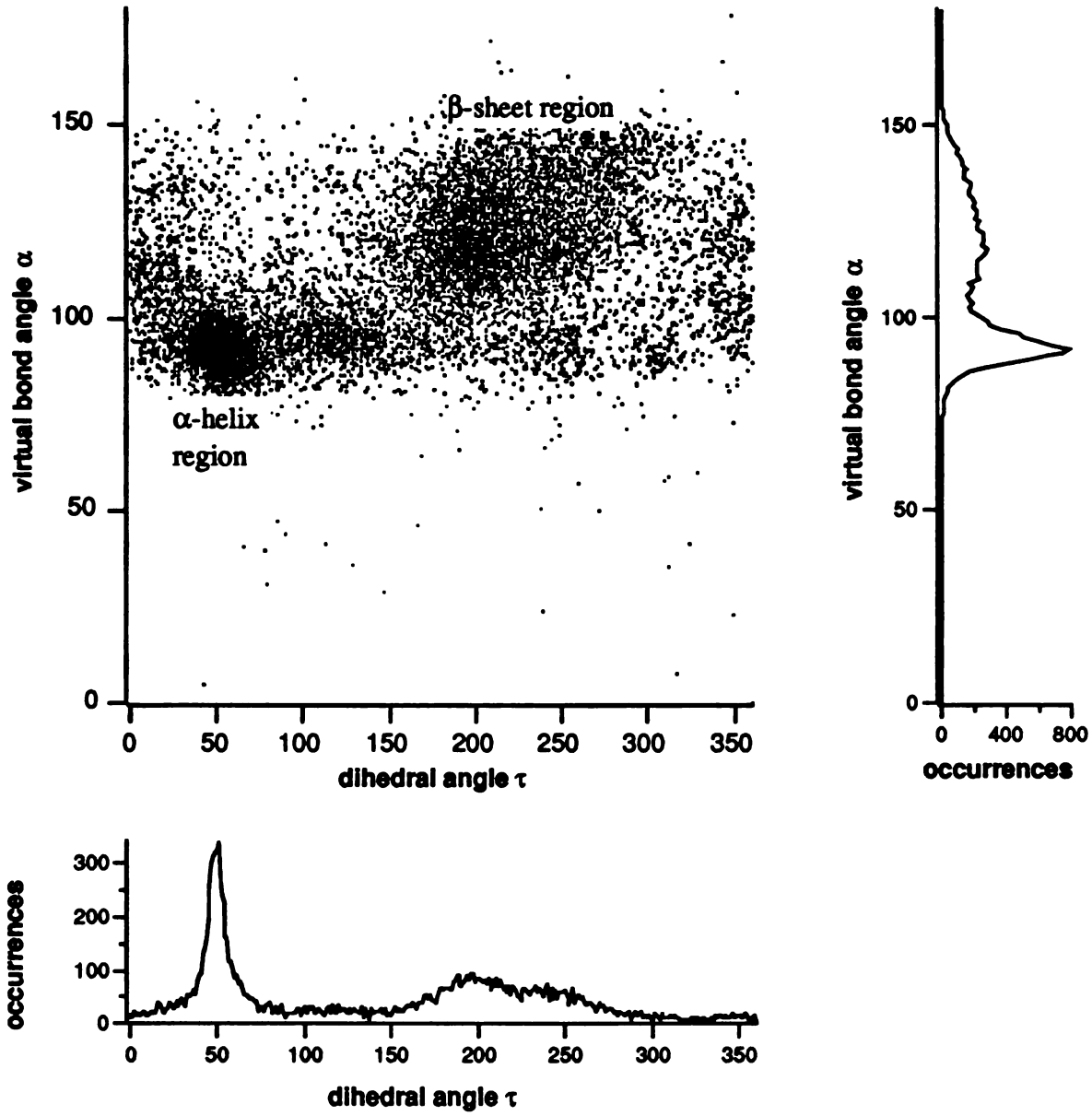


Figure 3.2. Correlated distribution of virtual bond angles α and dihedral angles τ in a set of 72 crytallographically-determined structures (Gregoret & Cohen, 1990). Distribution for α is shown at left and the distribution for τ is shown at bottom.

attempts, the previous atom is replaced. A maximum of 10^7 replacements is allowed before the program prints the atoms which it was able to place and exits with a time-out error.

Compactness is enforced because each model is constrained to be within an ellipsoid whose shape and dimensions are controllable. The size of the ellipsoid is scaled relative to the radius, R, a protein of N residues would have if it were perfectly spherical (Cohen & Sternberg, 1980; Gregoret & Cohen, 1990):

$$(2) \quad R = \sqrt[3]{\frac{3 \cdot N \cdot m}{4\pi \cdot \rho_P \cdot 10^6 \cdot N_A}} \cdot 10^{10}$$

(where N is the number of residues, m is the average molecular weight of an amino acid (110 g/mol), ρ_P is the density of globular proteins (1.4 g/ml, Creighton, 1983, p. 268), N_A is Avogadro's number, and 10^6 ml/m^3 and 10^{10} Å/m are unit conversion factors) Compactness is varied by scaling the volume in which random walk takes place by a factor ε:

$$(3) \quad V = \varepsilon \cdot \frac{4\pi R^3}{3} = \varepsilon \cdot \frac{4\pi abc}{3}$$

(where a, b, and c are the scaled ellipsoid vertices)

The program for generating random three-dimensional structures, RANPROT, was written in C. CPU time required for generating a structure depends both on the number of residues to be modeled and the volume constraint. Average time for generating a compact (scale factor $\varepsilon = 1.0$), native-like 138 residue protein is 0.3 minutes on a Silicon Graphics Personal Iris 20G. Average time for a 275 residue protein is 17.4 minutes.

In addition to these "native-like" random walk structures (incorporating native-like bond and torsion angles), we also produced structures on two types of lattices.

Cubic lattice random walk structures were generated using the same recursive algorithm, but restricting angles α to 0°, ±90°, and 180°. We also generated "knight's walk" random walk structures (Skolnick & Kolinski, 1990) on a cubic lattice taking steps like a knight's move on a chess board -- two steps in any direction on the lattice followed by a 90° turn and another step. The spacing of this lattice is 1.69Å so that inter-C^{α} distances are 3.8 Å. The excluded volume of each atom consists of that point itself and the six nearest lattice points. From any given residue i, the next residue in sequence, *i+1*, may be placed any of 24 places, excluding the position of residue i-1. Consequently, 10 different virtual bond angles α are possible and 58 different torsion angles τ are possible. Some bond angles are prohibited in the knight's walk structures: an angle of 180°, although possible on such a lattice, is not allowed since in true protein structures, the *i* to *i+3* inter-C^{α} distance is always less than 10.5 Å (J. Skolnick, personal communication). Similarly, very acute angles are excluded. Parameters and features of all three types of models are shown in Table III.1.

2. Construction of model sets

Several sets of structures were generated (Table III.2). We elected to simulate a 64 residue protein for ease of comparison with maximally-compact cubic lattice structures and to permit enough structures to be generated to ensure reliable statistics. Eight densities were studied for the native-like random walk model. One thousand structures were generated at each density, except for the most compact densities where 250 structures were made due to time constraints. Sets of structures at nine different densities were produced for the cubic lattice and knight's walk lattice. Apart from the maximally compact cubic structures which were constrained to 4x4x4 cubes, all structures were constrained to

Table III.1.

Properties of lattice and non-lattice protein chains

Property *Chain Geometry:*	**Real proteins**	**Native-like random walk structures**	**Cubic Lattice structures**	**Knight's walk lattice structures**
virtual bond length	3.8 Å	3.8 Å (adjustable)	3.8 Å (adjustable)	3.8 Å (adjustable)
virtual bond angle α	effective range 75° to 150°	0° to 180° in 2.5° increments (same effective range as real proteins)	90° or 180°	7 allowed angles: 66°, 78°, 90°, 102°, 114°, 127°, 143°
virtual dihedral angle τ	0° to 360° continuous range	0° to 360° in 2.5° increments	0°, 90°, 180°, 270°	0° to 360° (58 discrete angles)
α/τ correlation	correlated	correlated	uncorrelated	correlated
α/τ symmetry (about α = 90° and τ = 180°)	asymmetric	asymmetric	symmetric	symmetric
Packing Characteristics:				
closest observed or allowed contact	≈ 3.7 Å	3.75 Å	3.8 Å	2.4 Å
per-residue packing volume observed on average*, allowed†, or achieved‡	130 $Å^3$/Cα *	100 $Å^3$/Cα ‡	56 $Å^3$/Cα †	28 $Å^3$/Cα †
packing uniformity of maximally compact structures	generally uniformly compact	uniformity not guaranteed	uniformly compact	uniformity not guaranteed in our simulations
separation between i^{th} and $i+3^{rd}$ α-carbons in a helix	≈ 5.0 Å	≈ 5.0 Å	3.8, 5.4, 6.6, or 8.5 Å	2.4 to 5.5 Å
separation between adjacent α-carbons in sheets	≈ 5.1 Å	3.75 to 5.5 Å	3.8 Å	2.4, 3.4 or 3.8 Å

Table III.2. Compactness parameters of 64-residue random walk models

random walk type	volume scale factor (ε)	number of structures produced	mean radius of gyration (Å)	mean number of contacts
native-like	0.7	250	8.8 ± 0.1	67.6 ± 8.1
native-like	0.8	600	9.1 ± 0.2	59.6 ± 7.9
native-like	0.9	1000	9.3 ± 0.2	52.9 ± 7.8
native-like	1.0	1000	9.5 ± 0.2	47.5 ± 7.9
native-like	1.3	1000	10.2 ± 0.3	38.4 ± 7.6
native-like	1.5	1000	10.5 ± 0.4	35.3 ± 7.7
native-like	2.0	1000	11.2 ± 0.5	31.7 ± 7.7
native-like	1000.	1000	17.4 ± 3.5	22.2 ± 7.6
cubic lattice	0.48 (4x4x4 cube)	250	7.3	81.0
cubic lattice	0.5	250	7.5 ± 0.1	65.1 ± 4.6
cubic lattice	0.7	1000	8.2 ± 0.3	47.4 ± 8.4
cubic lattice	0.9	1000	9.0 ± 0.2	31.0 ± 6.3
cubic lattice	1.0	1000	9.2 ± 0.3	28.1 ± 6.1
cubic lattice	1.3	1000	9.8 ± 0.4	24.7 ± 6.1
cubic lattice	1.5	1000	10.1 ± 0.4	23.9 ± 6.1
cubic lattice	2.0	1000	10.7 ± 0.7	22.9 ± 6.4
cubic lattice	1000.	1000	17.6 ± 3.7	7.8 ± 5.4
knight's walk lattice	0.3	1000	6.6 ± 0.1	100.9 ± 9.5
knight's walk lattice	0.5	1000	7.6 ± 0.2	52.1 ± 8.1
knight's walk lattice	0.7	1000	8.2 ± 0.3	39.1 ± 8.0
knight's walk lattice	0.9	1000	8.8 ± 0.4	32.6 ± 7.7
knight's walk lattice	1.0	1000	9.0 ±0.4	30.7 ± 7.9
knight's walk lattice	1.3	1000	9.6 ± 0.6	25.7 ± 7.6
knight's walk lattice	1.5	1000	9.9 ± 0.6	24.5 ± 7.4
knight's walk lattice	2.0	1000	10.6 ± 0.8	21.3 ± 7.2
knight's walk lattice	1000.	1000	15.5 ± 3.4	15.1 ± 7.3

Footnote for Table III.2: The volume scale factor is the fraction of the volume expected for a protein assuming a uniform density of 1.4 g/ml as described in the text. For a 64-residue protein, the expected volume is 8.38×10^3 Å^3. The expected radius of gyration for a 64-residue protein can be estimated assuming a linear relationship between the radius of gyration (r_g)and the number of residues (N_{res}). For 24 crystallographically-determined small proteins with less than 150 residues, this relationship is linear and is given by $r_g = 0.042 \cdot N_{res} + 839$ ($r = 0.92$). A 64-residue protein would be expected to have $r_g = 11.0 \pm 0.7$ Å. The number of interresidue contacts is computed using a cutoff distance of 3.8 Å for the cubic lattice and knight's walk lattice structures and 5.5 Å for the native-like structures. Connected neighbors are not included as contacts. The ratio between the number of interresidue contacts (N_{cont}) and number of residues (N_{res}) for a set of 72 crystallographically-determined structures, is 0.75 ± 0.10 allowed to adopt any "reasonable" angle between 75° and 150° with equal probability, and τ any angle between 0° and 360° with equal probability. Eight sets of structures of the same densities as above were generated using this uniform distribution.

spheres. Table III.2 contains the average radii of gyration and the average number of interresidue contacts for the various sets of structures.

We studied the effect of chain length on the native-like random walks simulations using 58, 138, and 275 residue chains. Structures were generated at four different densities: ε = 1.0, 1.3, 1.5 and 1000 (effectively unconstrained). Since real proteins are generally slightly non-spherical, a prolate ellipsoid with an axial ratio of 1.25:1:1 was used to constrain these structures. No timeouts occurred during the generation of these nine sets of structures. To investigate the dependence on chain length in cubic lattice structures, one set of 250 maximally compact cubic lattice structures (i.e. restricted to a 3x3x3 cube) with a chain length of 27 was generated.

It should be noted that when we construct our model sets, we have no way of ensuring that we do not regenerate the same structure twice. To avoid this potential problem, a backtracking procedure, such as the scanning method described by Meirovitch and Lim (1990) should be used. Practically speaking, for the chain lengths used here, the chances of creating the same structure twice are infinitessimal. For a chain length of 64, even if each atom has a conservative choice of two conformations, there are $2^{64} \approx 1.85 \times 10^{19}$ total chain conformations possible. In the native-like structures, the number of possible conformations per residue is much larger than two, so the probability of reforming the same structure twice is negligible.

Native proteins are made from *L*-amino acids, and therefore adopt main chain torsional angles which minimize steric hindrance between the side chain at the α-carbon and the main chain. The lattice models, on the other hand, are achiral. In order to study the effect of chirality in our native-like random walk structures, we made the α/τ probability distribution, $p(\alpha,\tau)$, symmetric about $\tau = 180°$. The new probabilities, $p'(\alpha,\tau)$, at $\tau = x°$ and $\tau = 360 - x°$ were set to $\frac{1}{2}(p(\alpha, x) + p(\alpha, 360\text{-}x))$. Eight sets of structures with compactness ranging from ε = 0.70 to 1000 were generated using α,τ pairs from this new distribution. In addition, a uniform distribution without angular

preferences was made such that α is allowed to adopt any "reasonable" angle between 75° and 150° with equal probability, and τ any angle between 0° and 360° with equal probability. If in an all-backbone-atom model, ϕ and ψ were allowed to adopt any angle betwen 0° and 360°, but bond angles and bond lengths were held fixed, this mapping to α/τ space would result (J. Troyer, personal communication). Eight sets of structures of the same densities as above were generated using this uniform distribution.

We also made random structures constrained by ellipsoids of unusual shapes in order to identify other constraints which may influence secondary structure formation. Discus-shaped flavodoxin structures were made with ellipsoid vertex ratios of 2:2:1, 4:4:1, 10:10:1 and 15:15:1. Cigar-shaped structures were generated with vertex ratios of 2:1:1, 10:1:1 and 100:1:1. The volumes of all ellipsoids were normal (i.e. scaled by $\varepsilon = 1.0$).

RANPROT also has the capability of generating structures with correct secondary structure as listed in the HELIX and SHEET records of Protein Data Bank (PDB) files (Abola et al., 1987; Bernstein et al., 1977). Helical residues are constrained to angles $\alpha = 92°$ and $\tau = 50°$ whereas beta sheet residues are assigned angles $\alpha = 120°$ and $\tau = 200°$. Residues at the ends of helices and strands are allowed to adopt any conformation in the α/τ map. We were able to generate twenty-five 138-residue structures with the secondary structure observed in flavodoxin (Marquart et al., 1983) at a volume 130% of the expected native volume. Ellipsoid vertex ratios of 1.27:1.27:1 were chosen to correspond with the lengths of the eigenvectors of the principal moments of inertia of flavodoxin.

3. Evaluation of secondary structure content in models

Two methods were used to evaluate secondary structure in the random models, the difference distance matrix method of Richards and Kundrot (1988), and the con-

tact-based method of Chan and Dill (1990a; 1990b) Both methods use internal distances to assign secondary structure. The Chan-Dill method is attractive because it is theoretically applicable to any type of chain representation. It is the only useful method for cubic lattice and knight's walk lattice random walk structures since the Richards-Kundrot method is parameterized for real proteins. The Richards-Kundrot method is well-suited to these calculations because it requires only Cα positions.

A. Difference distance matrix definition

The method of Richards and Kundrot (1988) assigns secondary structure using a difference distance matrix. All inter-Cα distances in the protein are computed. These distances are then compared to the interresidue distances in an idealized segment of secondary structure. If the distances fall within some rmsd (root-mean-square deviation, Δr) limit, an assignment is made. We have found that for crystallographically-determined structures, the optimal value for helical assignment is $\Delta r = 0.75$ Å and the optimal value for strand assignment is $\Delta r = 0.5$ Å (N. Colloc'h and S. Presnell, personal communication). Using this method, Richards and Kundrot are able to identify five different types of secondary structure. Here we will only consider helices (α and 3_{10}) together, and β-strands. This method could theoretically be used for lattice structures, however, the Δr limits would have to be set to extremely large values and the accuracy of the method would be compromised.

B.Contact-based definition

The method of Chan and Dill (1990a; 1990b) is very similar to the Richards and Kundrot method, except instead of requiring specific, consistent interresidue distances, the method is only sensitive to whether a subset of inter-Cα distances is within some threshold value. Helices are identified if at least one of four sets of interresidue distance constraints is met (see Figure 3.3). A "type I" helix is defined by

contacts between alpha-carbons *i* and *i+3*, *i* and *i+5*, *i+1* and *i+6*, *i+2* and *i+7*, *i+4* and *i+7*; a "type II" helix is defined by contacts between alpha-carbons *i* and *i+3*, *i* and *i+7*, *i+1* and *i+6*, *i+2* and *i+5*, *i+4* and *i+7*; a "type III" helix is defined by contacts between residues *i* and *i+3* and *i+2* and *i+5*; and a "type IV" helix is defined by the contacts of a type II helix plus a contact between residues *i+4* and *i+7*. Residues assigned to more than one type of helical structure are counted only once.

The Chan-Dill definitions are entirely contact-based in two dimensions on the square lattice, but helices III and IV require coordinate information in three dimensions to completely specify chain geometry (H.S. Chan, personal communication). Here, since we apply the Chan-Dill definition to several types of structures, we use topological contacts only. The consequences are that on a cubic lattice, a type III helix is allowed to be "bent" rather than planar, much like a three-quarter length type IV helix. Similarly, a type IV helix is allowed to be planar (Figure 3.3).

Antiparallel sheets are defined if contacts between residues *i* and *j+2*, *i+1* and *j+1*, *i+2* and j occur simultaneously (for $j \geq i+3$). If $i+3 = j$, the antiparallel sheet occurs at a turn and residues *i+2* and *j* are counted as being in a turn while residues *i*, *i+1*, *j*, and *j+1* are assigned to antiparallel sheet conformation. Parallel sheets are defined if contacts occur between residues *i* and *j*, *i+1* and *j+1*, *i+2* and *j+2*. Chan and Dill require sheets to be planar on the lattice. Planarity is approximated here by requiring the virtual bond angles α_{i+1} and α_{j+1} to be greater than or equal to 120°. Unlike the Richards-Kundrot method, the Chan-Dill method has no definition for an isolated strand.

A distance cutoff of 5.5 Å was used to assign contacts in the random walk structures. For the cubic lattice structures, a distance cutoff of 3.8 Å was used, since Chan and Dill consider only the 6 nearest lattice points in tabulating topological contacts. The cutoff used in the knight's walk structures was also 3.8 Å. For this lattice type, each point is surrounded by 24 possible neighbor sites within at least 3.8 Å.

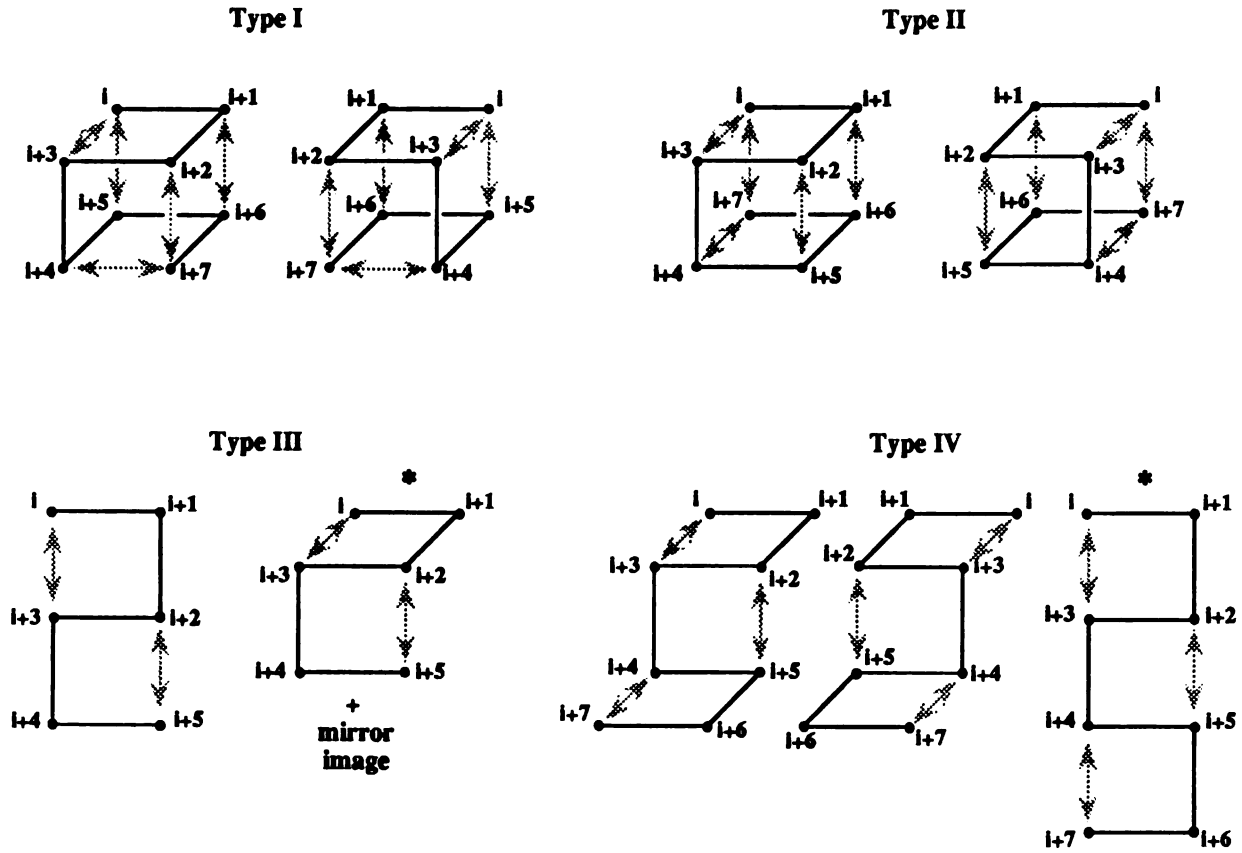


Figure 3.3. Lattice representations of helices. Topological contacts are shown as grey arrows. Helix types I, II and IV are chiral, so both streoisomers are shown. Helix type III is defined by Chan and Dill to be planar on the cubic lattice (H.S. Chan. personal communication), but we also allow the bent variety (*). Helix type IV, which is normally three-dimensional, is allowed to be planar (*) in this implementation of the Chan-Dill definition.

Results and Discussion

1. Examples of Random Walk Models

We carefully chose chain characteristics in order to develop the most protein-like random walk structures possible. Table III.1 shows a comparison of properties for real proteins, native-like random walk structures, structures restricted to cubic lattices and knight's walk lattice structures. Three random walk structures (native-like, cubic lattice and knight's walk) are shown at the same scale in Figure 3.4 along with the crystallographically-determined structure of pancreatic trypsin inhibitor (Marquart et al., 1983), a 58-residue protein. Helices and sheets, as identified by the contact-based method of Chan and Dill, are highlighted in red and blue respectively. The native-like structure shown is of a density comparable to real proteins. The cubic lattice structure shown is maximally-compact and the knight's walk lattice structure is as compact as we are able to generate ($\varepsilon = 0.3$).

The native-like structures can look very protein-like: packing is relatively efficient, secondary structure is observed, and reversals of chain direction are common near the surface. Apart from the large amount of random coil structure, there are several notable difference between these structures and real proteins. At densities greater than those of real proteins ($\varepsilon \leq 1.0$), the structures appear suspiciously spherical or ellipsoidal compared to real proteins (i.e. the fact that the walks were constrained to spheres or ellipsoids is obvious). Structures which are less dense than real proteins often appear nonuniformly packed -- one end of the molecule may be quite dense, while an unpacked loop may meander across to the other end of the molecule.

2. Secondary Structure Definitions

The Chan-Dill contact-based definition of secondary structure works remarkably well in identifying secondary structure in real proteins. The cutoff distance for

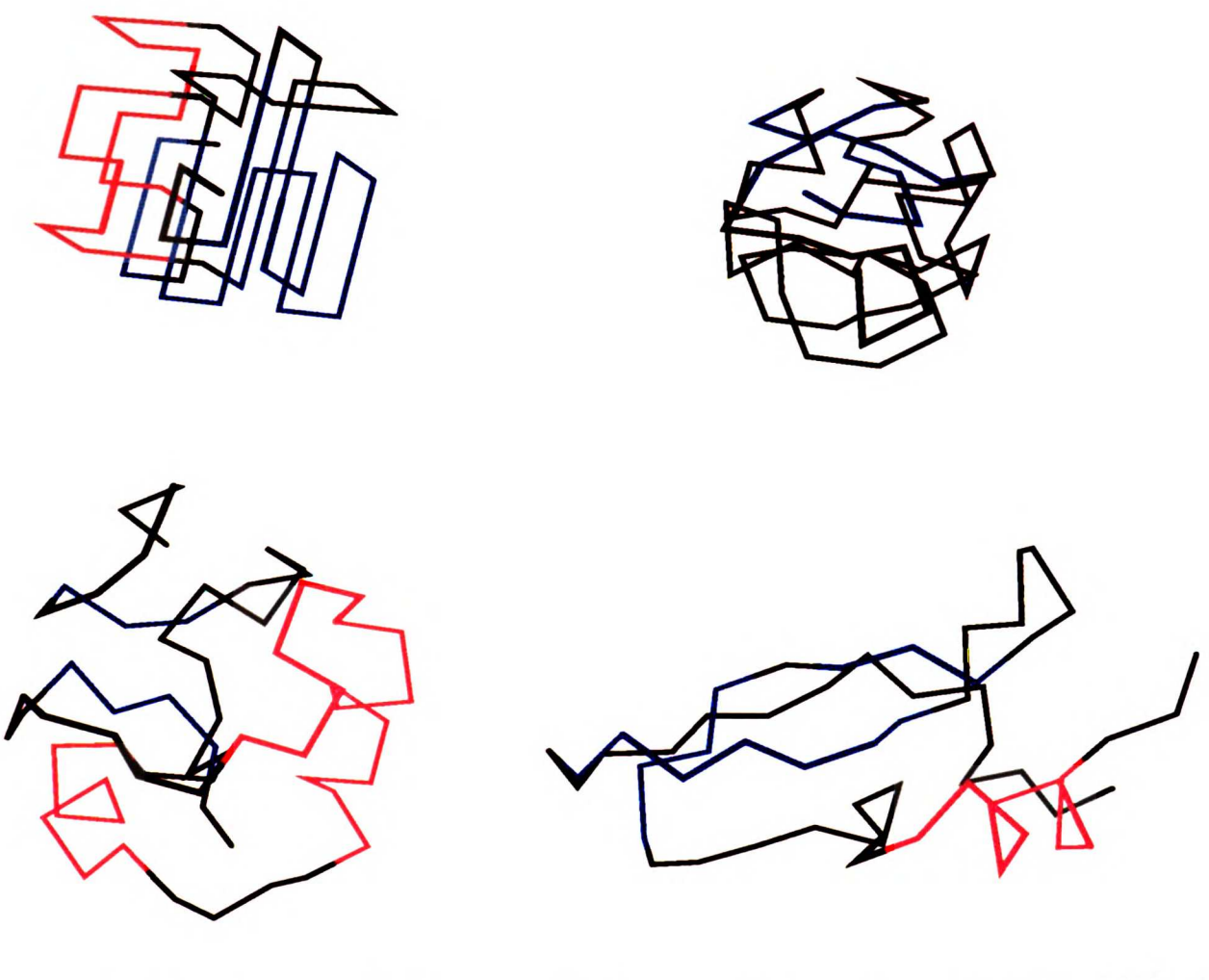


Figure 3.4. Several random walk structures and a crystallographically-determined structure for comparison. Top left: maximally compact 64-residue cubic lattice structure; top right: compact, 64-residue knight's walk lattice structure; bottom right: crystallographically-determined bovine pancreatic trypsin inhibitor structure (Marquart, *et al.*, 1983); bottom left: native-like 64-residue random walk structure. Helices are colored red and sheets are colored blue. Secondary structure was identified using the contact-based method of Chan and Dill. A cutoff distance of 3.8 Å was used to identify structure in the cubic and knight's walk lattice structures and a cutoff distance of 5.5 Å was used for the native-like and crystallographically-determined structures.

assigning contacts of 5.5 Å did not consistently over-assign or under-assign secondary structure in crystallographically-determined structures. Assignments very closely matched those in the HELIX and SHEET records of the PDB files. Using the contact-based method, 30 percent of residues in the data set of the 72 crystallographically-determined structures (Gregoret & Cohen, 1990) were assigned to helical structure and 34% were assigned to sheet structure (11% parallel and 25% antiparallel). These estimates of total secondary structure content are similar to those found using other methods (Chan & Dill, 1990b) although sheet content is slightly higher. The contact-based method will locate shorter, two-strand sheets since only six residues are required. Also, a hydrogen bond-based method, such as Kabsch-Sander (Kabsch & Sander, 1983) is likely to be more stringent because more atoms are required to make an assignment. Helix types I and II did not occur at all in the data set of real proteins. Interestingly, types I and II are rarely observed in cubic lattice structures as well. The large number of distance constraints required for these subtypes may select against them in the lattice models (Chan & Dill, 1990b). In addition, stereochemical constraints may prevent them from occurring in real proteins.

3. Compactness and Secondary Structure

A. Helical Structure

Figure 3.5A shows the dependence of helical content on compactness for the three classes of structures: native-like structures, cubic lattice structures and knight's walk lattice structures. For all structural classes, helical content increases as the mean radius of gyration decreases. As compactness constraints are removed, however, the amounts of helix in the various models differs significantly. The native-like proteins have a large baseline helicity. That is, unconstrained structures have 13% of residues in a helical conformation on average. The most compact native-like random walk structures which we could generate ($\varepsilon = 0.70$) contain an average of 28% helix. When

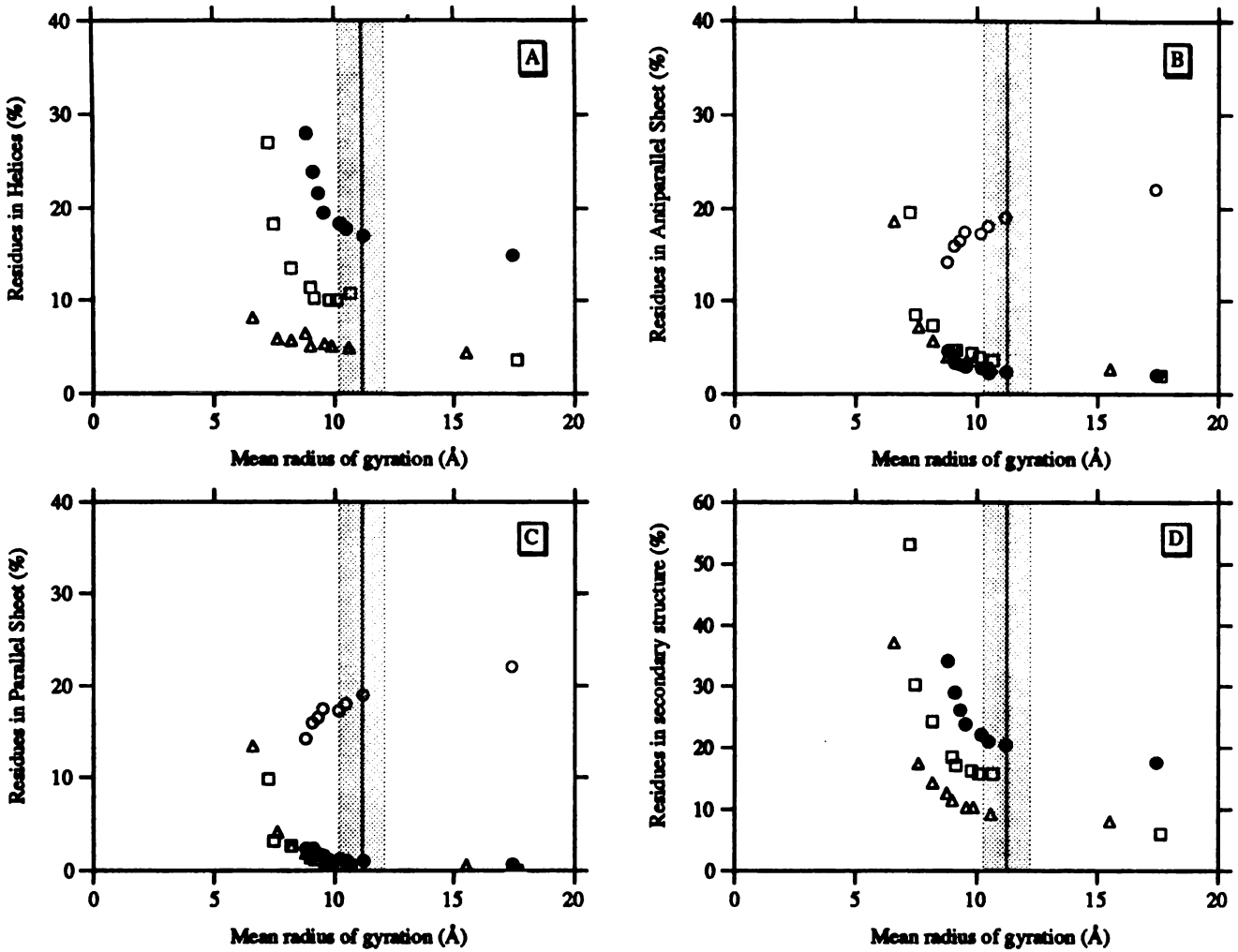


Figure 3.5. Secondary structure content in 64-residue random walk cubic lattice structures (❑), knight's walk lattice structures (Δ), and native-like random walk structures (●) as a function of mean radius of gyration (see table 2.) The expected radius of gyration for a 64-residue real protein is shown as a vertical line at 11.0 Å; ± one standard deviation (0.7 Å) is shaded in grey. **A.** Helical content. **B.** Antiparallel sheet content. Strand content in native-like random walk structures defined using the method of Richards and Kundrot is shown by open circles (O). **C.** Parallel sheet content. Strand content in native-like random walk structures defined using the method of Richards and Kundrot is shown by open circles (O). **D.** Total secondary structure content (helices and sheets.)

helical content in the native-like structures is defined using the method of Richards and Kundrot, the dependence on compactness is still observed. This suggests that the relationship between helical content and compactness is not an artifact of the contact-based definitions for helices which could result merely because the overall number of intrachain contacts is higher.

Maximally compact cubic lattice structures contain 27% helix on average. Unconstrained cubic lattice structures have a very low baseline helicity. This suggests that the increases in helical structure resulting from compactness is real and not a consequence of the limited number of virtual bond angles and torsional angles allowed on the cubic lattice. Because we are more lenient in the definitions of type III and IV helices and require only topological contacts rather than specific spatial geometry, our estimate of helical content is slightly higher than that of Chan and Dill (1990b) . Generation of a random sample of 27 residue maximally compact structures suggests that we overestimate helical content by about 6% when compared to Chan's and Dill's exhaustive study of maximally compact structures.

A dramatic increase in helicity is not observed in the knight's walk structures, although, the trend in the points on the graph suggests that at smaller radii of gyration, helical content may increase in these structures as well. Whether it will do so before the minimum radius of gyration is reached cannot be determined since we are limited by time constraints in generating more compact ($\varepsilon < 0.30$) structures. Nevertheless, it appears that the knight's walk lattice does not map as well to α-helical conformation. Different lattice models may have different inherent compact subconformations. Instead of helices, the knight's walk lattice structures may have an alternative type of regular repeating structure not observed in native proteins.

In native proteins, there is a linear relationship between the number of amino acids and the radius of gyration (see legend, Table III.2). The radius expected for a 64 residue protein is indicated in Figure 3.5. Real proteins are not as compact as the

most compact native-like structures we are able to generate nor as the maximally compact cubic lattice models when judged by both the radius of gyration and also by the number of intrachain contacts (see Table III.2).

A possible bias in our structure generation method is that it is unidirectional: we generate structures from the amino terminus and add residues sequentially. Residues at the beginning of the chain are likely to be subject to fewer excluded volume constraints. Indeed, very compact native-like random walk structures ($\varepsilon < 1.0$) contain 40-60% more residues in a helical conformation at their C-termini than at their N-termini. This result shows that our method of conformational searching does not always produce a truly random set of structures as an exhaustive search would. Structures with a compactness comparable to proteins, however, are not biased in this manner.

B. Sheet Structure

All three types of structures exhibit a similar dependence of sheet content on the radius of gyration (Figures 3.5B and C). The native-like random walk structures contain little sheet structure. It appears as though sheet content would increase if we were able to generate more compact structures. However, the increase in sheet content is accompanied by a drop in total strand content (open circles; computed using the Richards-Kundrot method), so it is unlikely that the native-like structures will ever have as many sheets as cubic lattice structures.

For the cubic lattice structures, a significant amount of sheet structure is seen only when maximal compactness is achieved. Compact knight's walk lattice structures also contain a significant amount of sheet structure. The preponderance of sheet structure in both types of compact lattice structures is likely to come from a bias in the lattice, which contains ordered arrays of points on which to form sheets.

Figures 3.5D shows total secondary structure content as a function of the radius of gyration. Maximally compact cubic lattice structures have the most secondary struc-

ture, to which both helix and sheet contribute equally. Secondary structure content in the native-like structures is largely dominated by helical structure. Compact knight's walk structures are dominated by sheet structure.

4. Dependence on chain length

Secondary structure content in the native-like random walk proteins depends little on chain length. Figure 3.6 shows secondary structure content in compact and unconstrained 58, 138, and 275 residue structures as evaluated by both secondary structure assignment methods. Although there is more variation in the amount of secondary structure with the shorter chains, the average secondary structure content is independent of length. There is also little difference in secondary structure content between the compact ($\varepsilon = 1.0$) structures and unconstrained structures: helical and sheet content drops slightly while strand content
increases. Presumably this is because the chain can take longer excursions in a given direction before encountering either itself or the ellipsoid boundary.

The effect of compactness on the overall distribution of secondary structure content in a set of structures (e.g., 138-residue structures) is shown in Figures 3.7A through D. Helical content accounts for most of the secondary structure content. Peaks at 6, 12, and 18 residues reflect the minimum number of residues required to specify a helix or a sheet. These peaks become more pronounced when the compactness constraint is removed. This occurs because there is more variation in helix (or sheet) length in more compact structures. Less compact structures are likely to have one, two, or three isolated helices (or sheets). Interestingly, structures constrained to ellipsoids 30% larger than the expected volume of real proteins still have a distribution which resembles that of the more compact structures. This suggests that slightly expanded structures, perhaps corresponding to a molten globule-like state (Kuwajima, 1989), may share some features with more compact, folded structures.

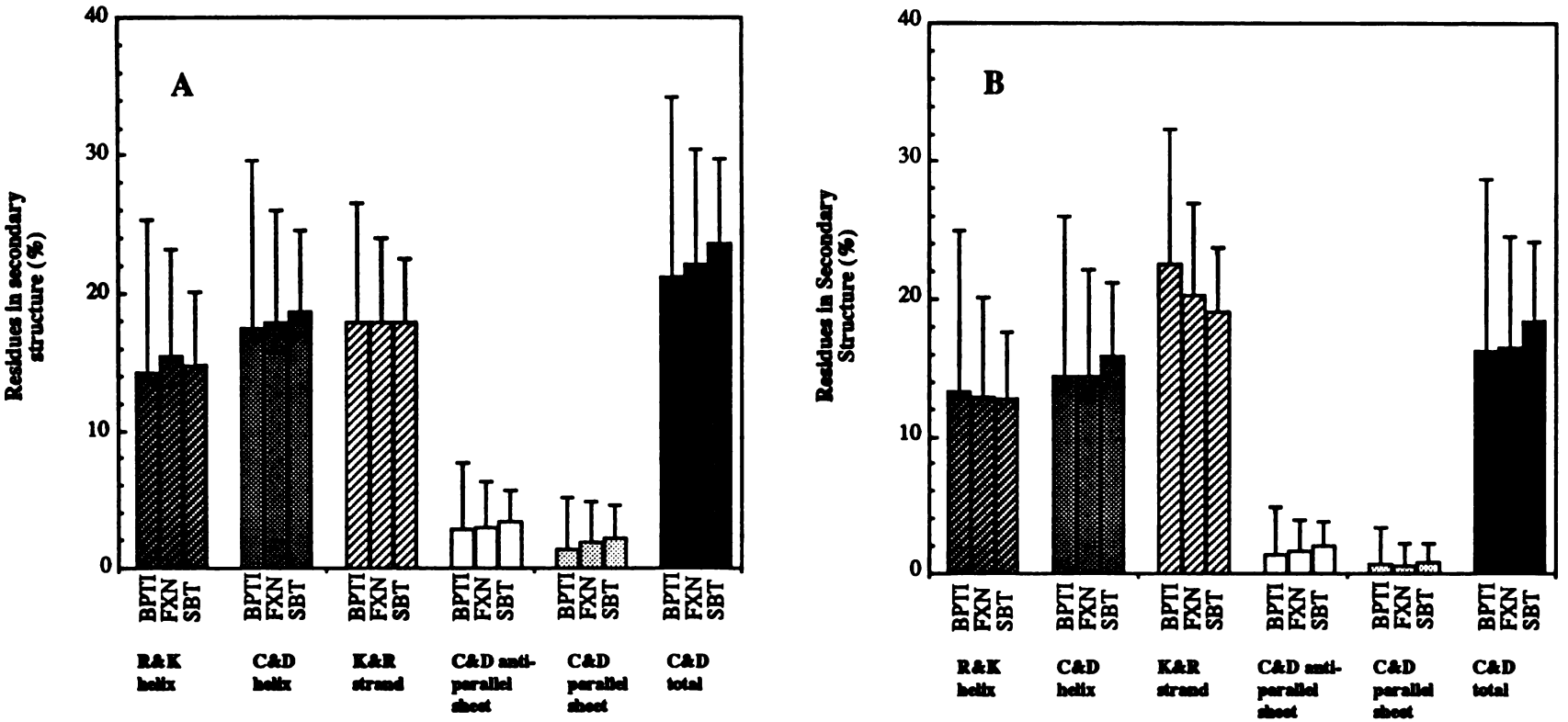
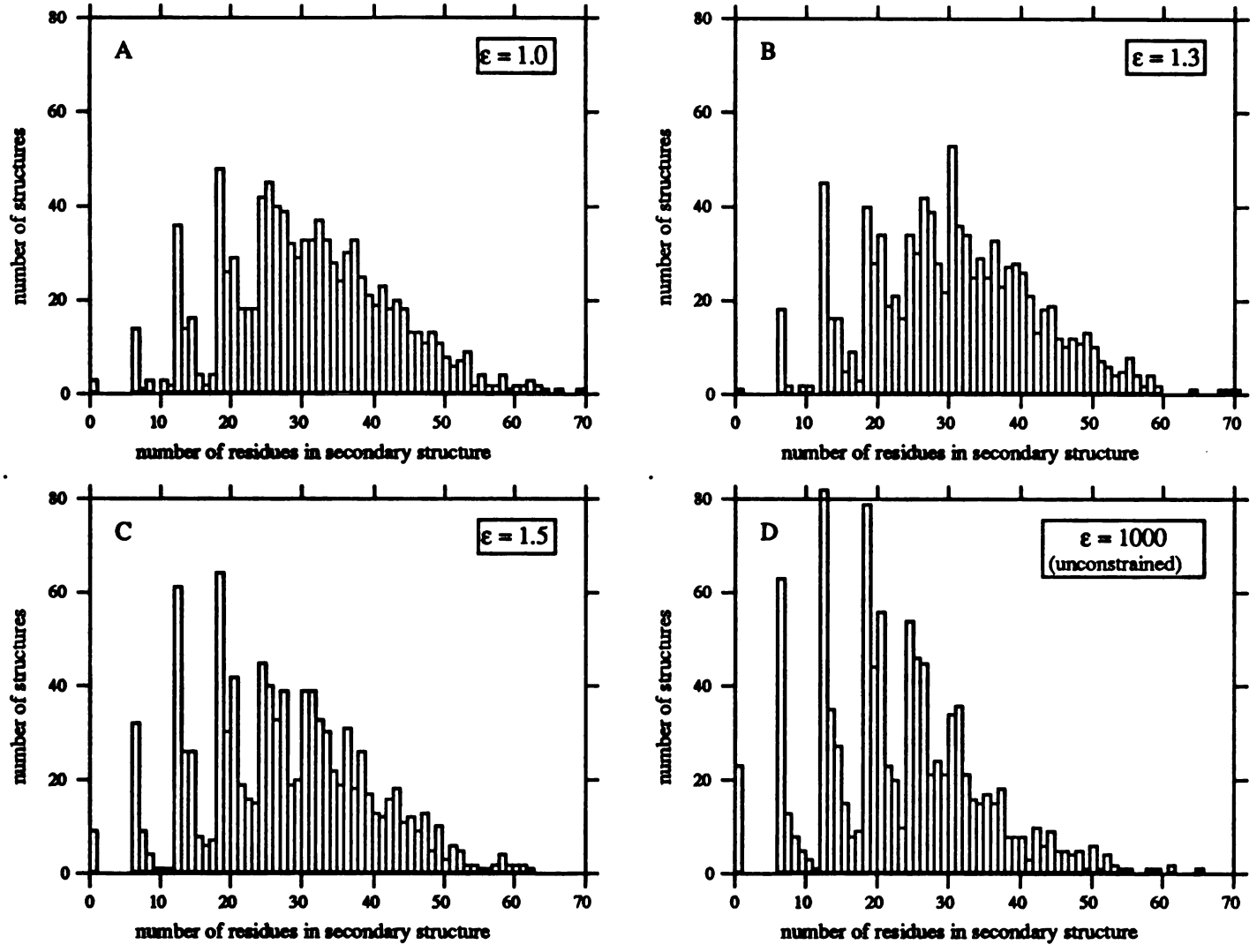


Figure 3.6A. Secondary structure content in native-like 58-residue (BPTI), 138-residue (FXN) and 275-residue (SBT) structures of a density comparable to native proteins: dependence of secondary structure content on chain length. Abbreviations used: C&D: Chan-Dill method, R&K: Richards-Kundrot method. **B.** Secondary sturcture content in unconstrained native-like BPTI's FXN's and SBT's.



Figures 3.7A-D. Histograms showing total secondary structure content in native-like random walk flavodoxins of different densities. The volume scale factor is given at top right. Secondary structure was evaluated using the Chan-Dill method.

The work of Chan and Dill (1990b) focused only on short chains to allow exhaustive sampling of conformation space. Since computation time for a full search of conformation space increases as C^N, where C is the coordination number of the lattice and N is the chain length, searches for much longer sequences were prohibitive. The average secondary structure content in maximally compact 27-length proteins was 50%. We found that we could duplicate the observed distribution of secondary structure in maximally compact 27-length structures with a random sample of 250 walks on cubic lattices (restricted to a 3x3x3 cube). In addition, with a random sample of 64-length structures, total secondary structure remained at 50%. This suggests 1) that the results of Chan and Dill are applicable to longer chains, and 2) a random sample of compact conformations can give similar results to an exhaustive conformational search. In two dimensions, Chan and Dill postulate that secondary structure content approaches 100% in maximally compact chains of infinite length. In three dimensions, it appears that the limiting value may be closer to 50%.

5. Effect of chirality

Helical content may be over-estimated in the cubic lattice studies because chirality is absent. To see what effect the asymmetric α/τ probability distribution has on secondary structure, we made the distribution symmetric to allow right- and left-handed helices and strands to form with equal probability. We found a significant decrease in the overall amount of secondary structure at all densities. The helical content of unconstrained structures is only 7%, as compared to 15% in unconstrained structures generated using the asymmetric α/τ distribution (Figure 3.8A). In the most compact structures ($\varepsilon = 0.70$), helical content is only 13% -- half as much as in structures of equal density generated from the asymmetric α/τ distribution. This result is surprising since it may be expected that allowing right-handed, left-handed, and planar

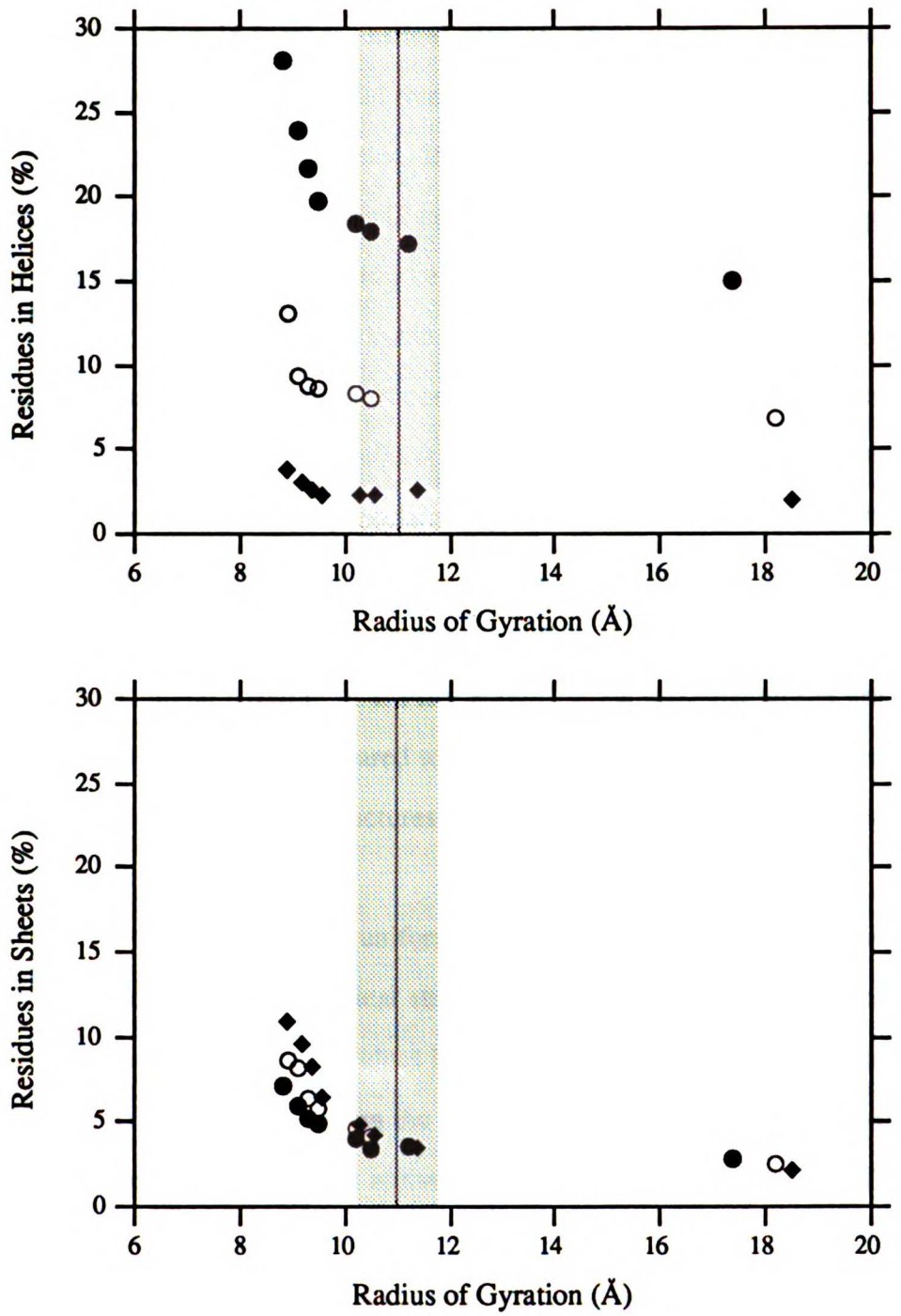


Figure 3.8A. Helical content in native-like structures (●), native-like structures made using a symmetric α/τ distribution (O), and structures generated using a flat distribution (♦). B. Sheet content (antiparallel + parallel).

(achiral) helices would *increase* helical content. However, the opposite occurs since the chain has more conformational freedom. Consider four connected residues on their way to forming a right-handed helix. If the fifth residue has equal probability of continuing in a conformation that would extend the right handed helix or adopting a left-handed conformation, the helix may be broken. Perhaps this is the origin of glycine's helix breaking behavior. On the cubic lattice, conformation space is still restricted enough that helices are observed in compact conformations. When the constraint of the lattice is removed and handedness is eliminated, very little regular structure is found. This phenomenon may explain why so little helical structure is observed on the knight's walk lattice. With respect to helices, the knight's walk lattice structures are more similar to the symmetric α/τ native-like structures because there is much more conformational freedom in τ. Fifty-eight discrete torsional angles are allowed on this lattice (Table III.1) as compared to only four on the cubic lattice. Sheet content in the symmetric torsion angle structures is about equal to that in the asymmetric native-like structures (Figure 3.8B).

When the α/τ distribution is uniform (i.e. there is no preference for particular virtual bond or torsion angles) helical structure nearly disappears. The contact-based definition detects only 4% of residues being in a helical conformation at the most compact density (Figure 3.8A). When the Richards-Kundrot definition is used, only 0.5% of residues are considered to be native-like helices in these structures. Presumably, as the overall number of interresidue contacts increases, the probability of forming contacts consistent with the Chan-Dill definitions of helices increases. Interestingly, the uniform distribution structures have slightly more sheet structure than the native-like structures -- 11% in the most compact structures as opposed to 7% in asymmetric structures of similar density (Figure 3.8B). The reason for this behavior is the larger area of the α/τ plot which is now accessible to strand conformation coupled with loss of the strong preference for helical structure.

6. Shape and Secondary Structure

In order to investigate other factors which may influence secondary structure content, we generated structures constrained to extremely flat and extremely oblong ellipsoids. Figure 3.9A shows helical content as measured by both the Richards-Kundrot and Chan-Dill methods. Shape has little effect on helix formation until very extremely prolate structures are generated. When the vertex ratio is 100:1:1, helical content increases dramatically. We observed many more program timeouts in the generation of the less extremely prolate proteins (e.g. with vertex ratios of 10:1:1) than in the more extreme case (100:1:1). This result suggests that there is some intermediate set of boundary conditions for which it is difficult to generate structures. Once this barrier is surmounted, the chain may have fewer choices of dihedral angles, but polymerizes quickly. This effect could have significance to the formation of filamentous proteins.

Strand and sheet content is influenced less by shape. Only the discus with vertex ratios of 15:15:1 enhances strand content (Figure 3.9B). We do not see a complementary increase in sheet content in these structures as we expected. As in the compact native-like random walks constrained to spheres or slightly off-spherical ellipsoids, strand-strand pairing remains a rare event. Figure 3.10 shows a cigar-shaped 138 residue chain constrained to an ellipsoid with vertex ratios of 100:1:1 and a discus-shaped 138 residue chain constrained to an ellipsoid with vertex ratios of 15:15:1.

7. Effect of imposing "correct" secondary structure

Structures with 138 residues and helices and strands distributed along the chain as seen in flavodoxin (Smith et al., 1977) were generated in order to investigate the effect of introducing local constraints on non-local interactions. We found that the mean rms deviation of a set of twenty-five random walk flavodoxin structures from the crystallographically-determined structure was 15± 1 Å. The same mean rms deviation

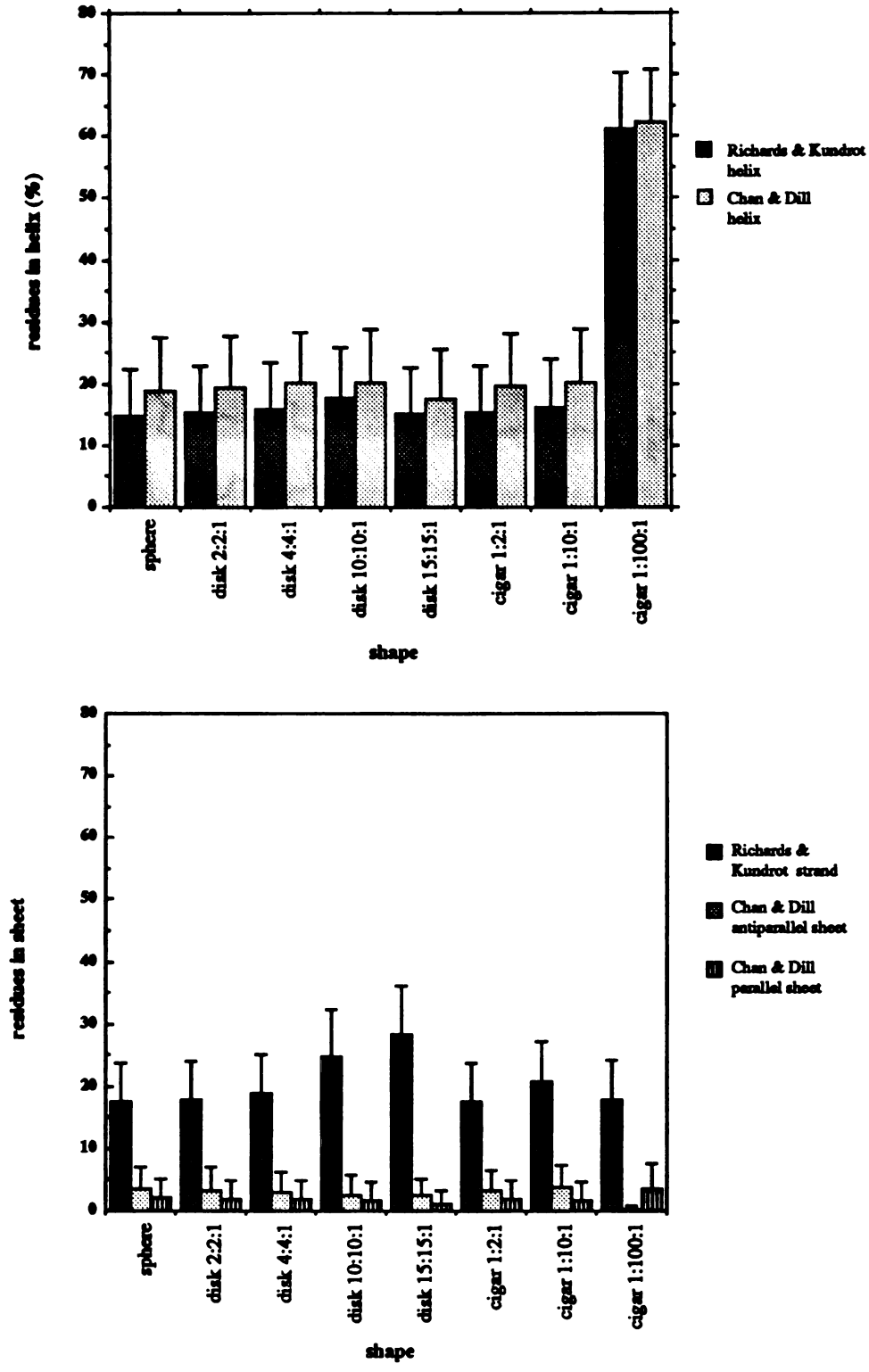


Figure 3.9A. Helical content in structures generated in extremely oblate (disk-shaped) and prolate (cigar-shaped) ellipsoids. **B.** Sheet and strand content.

Figure 3.10. Top: two views of a structure resulting from constraint by an ellipsoid with vertex ratio 15:15:1. Helices and sheets, as defined by the Chan-Dill method are colored red and blue respectively. Bottom: two views of a structure resulting from constraint by an ellipsoid with vertex ratio 100:1:1.

was found for the set of 1000 native-like random walk structures of equal density (ε = 1.3). Strand-strand pairing is as rare in these structures as in ordinary native-like random walk structures. Only 3% of residues were in antiparallel sheets and 2% in parallel sheets. These structures have volumes 30% larger than the expected volume of flavodoxin. Computer time constraints precluded the generation of more compact structures. It is possible that more compact conformations could have more sheet structures.

Conclusions

Compactness induces secondary structure formation in all three lattice and non-lattice representations of polypeptide chains studied here. The types of structures which predominate differ according to the type of representation used. Native-like random walk structures are predominantly helical. Random walks performed on a cubic lattice contain approximately equal amounts of helix and sheet and compact structures constructed on a knight's walk lattice have a significant amount of sheet structure but little helix.

A much larger amount of sheet structure is seen in compact cubic lattice structures and knight's walk lattice structures than in the non-lattice random walk structures. This leads us to believe that the lattice imposes a bias which favors sheet formation: the lattice contains ordered linear arrays of points upon which paired strands can lie. Our model lacks this bias. In our model, local structures, such as helices and strands form with a fairly high probability even in volumetrically unconstrained structures: 13% of residues, on average, will be in a helical conformation in unconstrained structures, and 20% will be in a strand conformation. We can reconcile the lack of sheet structure in our model with that obtained on the cubic lattice (Chan & Dill, 1990b) by suggesting that the linearly arranged lattice points acts as a guide for sheet formation, much as hydrogen bonds may "lock in" strand-strand pairing during the folding of real pro-

teins. The geometry of the cubic lattice model introduces an additional attractive term which promotes the formation of non-local interactions like sheets. If we had included an attractive term in our native-like random walk model, we may have seen sheets form as well.

Helical structure in the native-like random walk structures is halved when torsion angles of both handednesses are equally probable. Fewer helices form because the chain has more conformational freedom. The inclusion of right-handed, left-handed and achiral helices in the cubic lattice model probably does not overestimate helical structure. The lack of helical structure in the knight's walk lattice structures may result because torsion space is extensive yet symmetric. This is consistent with the tendency of glycine to terminate secondary structure in proteins. When the preference for helical structure is taken away by making the α/τ distribution uniform, helices vanish almost entirely.

Are maximally compact structures a valid reference state for real proteins? Structures which have high secondary structure content are much more compact than real proteins. In our simulations we see enhancement of secondary structure only in native-like random walk structures which have volumes 20-30% smaller than typical protein volumes. Though the cubic lattice may be scaled such that lattice points are the same distance apart as sequential alpha carbons in real proteins (3.8 Å), the density of alpha-carbons in the maximally compact state is twice as great (see Table III.1.) Efforts to map alpha-carbon positions of real proteins onto a cubic lattice has resulted in structures which have only 75% of the available lattice points filled (D. Yee, personal communication). It may not be relevant, therefore, to compare maximally compact structures to real proteins. Still, it could be argued that real proteins are maximally compact since they are as compact as close packed spheres (Richards, 1977). Because we neglect explicit side chains in our models, the excluded volume constraints may not take effect until densities greater than those in real

proteins are attained. However, studies of molten globule protein folding intermediates suggest that protein structures which are 10-30% larger in volume than the native state still contain a significant amount of secondary structure (Kuwajima, 1989). Therefore, we would have expected to have seen a greater amount of secondary structure at least at densities comparable to native proteins.

Constraints other than compactness can have effects on secondary structure formation. Extremely prolate ellipsoidal (cigar-shaped) structures have a greatly enhanced amount of helical structure. Discus shaped structures have more extended strands but these are not arranged to form beta-sheets. Apparently, the entropic cost of strand-strand pairing in our simulations is too great.

Compactness clearly drives the formation of compact substructures such as α-helices and β-sheets. We see greatly enhanced helical content in super-compact native-like random walk structures. Sheets occur less frequently in our native-like random walk models. Even though they may be an inherently compact conformation, compactness alone does not drive their formation. Clearly, lattices impact subtle conformational bias to chain simulations which can quantitatively influence conclusions about the origins of protein structure.

References for Chapter 3

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. (1987). In *Crystallographic Databases -- Information Content, Software Systems, Scientific Applications*. (Allen, F.H., Bergeroff, G. & Sievers, R., ed.), pp. 107-132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

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.

Chan, H.S. & Dill, K.A. (1989). Compact Polymers. *Macromolecules*. **22**, 4559-4573.

Chan, H.S. & Dill, K.A. (1990a). Effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys*. **92**, 3118-3135.

Chan, H.S. & Dill, K.A. (1990b). Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA*. **87**, 6388-6392.

Cohen, F.E. & Sternberg, M.J.E. (1980). On the Prediction of Protein Structure: The Significance of the Root-mean-square Deviation. *J. Mol. Biol*. **138**, 321-333.

Covell, D.G. & Jernigan, R.L. (1990). Conformations of Folded Proteins in Restricted Spaces. *Biochemistry*. **29**, 3287-3294.

Creighton, T.E. (1983). Proteins: Structures and Molecular Properties. New York, W. H. Freeman and Co.

Go, N. & Taketomi, H. (1978). Respective role of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. USA*. **75**, 559-563.

Gregoret, L.M. & Cohen, F.E. (1990). Novel Method for the Rapid Evaluation of Packing in Protein Structures. *J. Mol. biol*. **211**, 959-974.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded geometrical features. *Biopolymers*. **22**, 2577-2637.

Kuwajima, K. (1989). The Molten Globule State as a Clue for Understanding the Folding and Cooperativity of Globular-Protein Structure. *Proteins: Struct. Func. Genet*. **6**, 87-103.

Levitt, M. (1975). A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding. *J. Mol. Biol*. **104**, 56-107.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). The Geometry Of The Reactive Site And Of The Peptide Groups In Trypsin,

Trypsinogen And Its Complexes With Inhibitors. *Acta Crystallogr., Sect.B.* **39**, 480-490.

Meirovitch, H. & Lim, H.A. (1990). Computer simulation study of the θ-point in three dimensions. I. Self-avoiding walks on a simple cubic lattice. *J. Chem. Phys.* **92**, 5144-5154.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1988). Numerical Recipes in C: The Art of Scientific Computing. Cambridge, Cambridge University Press.

Ramachandran, G.N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95-99.

Ramachandran, G.N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Prot. Chem.* **23**, 283-437.

Richards, F.M. (1977). Areas, Volumes, Packing and Protein Structure. *Ann. Rev. Biophys. Bioeng.* **6**, 151-176.

Richards, F.M. & Kundrot, C.E. (1988). Identification of Structural Motifs From Protein Coordinate Date: Secondary Structure and First-Level Supersecondary Structure. *Proteins: Struct., Func., Genet.* **3**, 71-84.

Skolnick, J. & Kolinski, A. (1989). Computer simulations of globular protein folding and tertiary structure. *Ann. Rev. Phys. Chem.* **40**, 207-235.

Skolnick, J. & Kolinski, A. (1990). Simulations of the Folding of a Globular Protein. *Science.* **250**, 1121-1125.

Smith, W.W., Burnett, R.M., Darling, G.D. & Ludwig, M.L. (1977). Structure Of The Semiquinone Form Of Flavodoxin From Clostridium Mp: Extension Of 1.8 Angstroms Resolution And Some Comparisons With The Oxidized State. *J. Mol. Biol.* **117**, 195-225.

Chapter 4

Structure Prediction by Homology Modeling

Part A. Introduction to Modeling by Homology

Part B. Modeling the structure of the cercarial elastase from *Schistosoma mansoni*

Part C. Improving side chain positioning

Part A. Introduction to Modeling by Homology

The most accurate method for protein structure prediction is homology-based modeling. This method relies on sequence similarity between the protein of interest and a protein (or proteins) whose three-dimensional structure has been determined experimentally. The sequence of interest is first aligned with that of the protein whose structure has been determined. Then, following the alignment, the backbone of the known protein is fitted with the amino acid side chains of the protein of interest. Obviously, the availability of a structure with significant sequence similarity is a stringent requirement which limits the applicability of this technique. However, it has been speculated (Dorit *et al.*, 1990) that there is a limited number of folding motifs and that as the structural database grows, homology modeling will become the primary means for predicting protein structures. Homology modeling is most successful when the sequence identity between the protein of interest and the known structure is high (> 75%). Although there can be significant *structural* similarity between proteins having as little as 25% sequence similarity, proteins this dissimilar often have insertions and deletions, making homology modeling a challenge. Homology modeling, also known as "knowledge-based" protein structure prediction was recently reviewed by Blundell *et al.* (1987).

The first step in homology modeling is the generation of a sequence alignment. A correct alignment between the sequence of interest and the sequence whose structure is known is the most important step in the modeling process, since all subsequent modeling depends upon it: if the alignment is incorrect, the entire model is incorrect (Greer, 1990). It is often advantageous if the structures of several homologous proteins are available. A multiple, structure-based sequence alignment can then be generated by superimposing the known structures in space and noting which residues in the two (or more) structures occupy equivalent positions. The sequence of the protein of interest is then aligned either manually or automatically to this consensus alignment. Aligning the known structures in space is a complicated problem which is not easily automated because two sequences sharing structural identity may have

different numbers of residues and several insertions and deletions. Consequently, one must either predefine superimposable residues (thus defeating the purpose of automation) or automatically generate trial rotation/translation matrices and then iteratively check for quality of superposition. Such methods are currently under development at Birkbeck College (Sutcliffe *et al.*, 1987). Recent work by Taylor and Orenga (Taylor & Orenga, 1989) uses a dynamic programming method to superimpose structures by aligning their interresidue distance matrices, or contact maps. This method has the advantage that it does not require the user to predefine homologous regions.

It is important to recognize that a structure-based alignment is not necessarily the evolutionarily relevant one. Although it has been observed that regular secondary structure elements, such as helices and strands, are generally well conserved while insertions and deletions occur in loops (Chothia *et al.*, 1986), this is not a rule: intra-helical and intrastrand frameshifts resulting from an insertion can occur and have been shown to be tolerated when engineered (Sondek & Shortle, 1990). Also, the observation that different evolutionary trees are deduced using parsimony methods from sequences aligned by structural alignment and sequence alignment (Johnson *et al.*, 1990) suggests that the structure-based alignment is not necessarily evolutionarily correct. However, for the purpose of building a model based on known structures, a structure-based alignment is preferred.

Sequences of related, though structurally unsolved proteins are also useful, particularly when the sequence identity between the protein of interest and the known structure(s) is low. Confidence in the alignment in regions of low similarity can be gained if additional sequences are used to generate a multiple sequence alignment. The multiple sequence alignment method of Taylor (Taylor, 1986) uses information from structurally-aligned homologous proteins to generate sequence "templates" against which additional sequences are aligned. The method of Smith and Smith (1990), although primarily a tool for locating homologous members of the same protein family from a large database of sequences, makes all possible pairwise comparisons of

sequences and generates patterns common to both sequences. Additional sequences are accumulated in a tree pruning fashion if they also match the pattern. Both of these techniques can generate reasonable alignments from many sequences even when the individual pairwise sequence identity between set members is low.

Apart from sequence alignment, the two major unsolved problems of homology model-building are loop building and side chain modeling. Loops are difficult to model because this is where most insertions and deletions occur. Even in cases where loop length is preserved, the three-dimensional structure of the loop may not be constant across members of the same protein family, especially if sequence identity in this region is low. Most approaches to loop modeling have utilized loop libraries (Jones & Thirup, 1986; Kneller, 1988). A loop library is a set of peptide conformations derived from the crystallographic database of protein structures. The library is searched to find a conformation which has the correct number of residues and overlaps well with the backbone of the model where the loop is to be fused. The problem with this method is that both suitable and unsuitable loops may be found and sorting between them can be difficult. Chothia and co-workers (Chothia & Lesk, 1987; Chothia *et al.*, 1989) have successfully modeled the hypervariable loop regions of an immunoglobulin using a more limited set of loop conformations derived from a basis set of immunoglobulin structures. This method works well but depends on the availability of a large set of similar structures from which to derive the loops.

Bruccoleri and co-workers have applied a conformation searching technique to modeling the hypervariable loops in antibodies (Bruccoleri *et al.*, 1988). Their program, CONGEN, uses an angular grid search method. The coarseness of the grid is defined by the user and alternative conformations are evaluated using the CHARMM force field (Brooks *et al.*, 1983). For the two structures modeled, the average backbone r.m.s. (root-mean-square) deviations from the true structure were 1.4 Å and 1.7 Å. The major difficulty of this method is that it is time consuming and therefore limited to eight-residue loops. Longer loops must be split in two.

Modeling the conformations of side chains is difficult because of the number of degrees of conformational freedom involved. There are several ways in which the search may be minimized and the problem has mainly been approached in two ways -- using a rule-based approach and using a conformational searching approach. The rule-based approach focuses on the observation that side chains which differ in two homologous structures *usually* have the same torsional angles (Summers *et al.*, 1987; Sutcliffe *et al.*, 1987). Therefore, when making a substitution, a good first approximation may be to place the side chain of the model in a similar conformation as in the known structure or structures. If the side chain in the modeled protein is much larger than in the known structure (for example, Ala -> Phe), then the best guess may be to choose the most frequently observed rotamer as tabulated by Ponder and Richards from a set of highly-refined protein structures (Ponder & Richards, 1987). Further refinement of rotamer choice may be made by taking into account the type of secondary structure the in which the residue is located (McGregor *et al.*, 1987).

Several grid searching methods for the prediction of side chain conformations have been developed. Early attempts focused on modeling the conformations of only one or two amino acids (Shih *et al.*, 1985; Snow & Amzel, 1986). The success of these studies is hard to evaluate because of the small number of examples.

The CONGEN program of Bruccoleri *et al.* (1987; 1988) uses an angular grid search to model both the conformations of loops (see above) and the conformations of side chains. The grid search is adjustable with a maximum coarseness of 120° with minima at ±60° and 180° corresponding to the two *gauche* and the *trans* conformations. A finer search of 30° or 60° steps is usually employed. Side chains are added to a structure one at a time either in sequential order or from the center of mass out. The conformation space of each added side chain is searched and the lowest energy conformation is saved. This method is able to replace side chains onto a known structure with an r.m.s. deviation of 2.5 Å.

Recently, Schiffer *et al.* (1990) described a method, called LECS for lowest energy conformational searching, for predicting side chain conformations using molecular

mechanics. The conformation space of each amino acid in turn is searched: each rotamer (Ponder & Richards, 1987) of that amino acid is substituted and that residue and the surrounding region is subsequently energy-minimized using AMBER (Singh *et al.*, 1987; Weiner *et al.*, 1984). The rotamer with the lowest energy is chosen as the prediction. This method is computer-intensive, but seems to work well, particularly for core residues. This group is currently investigating incorporating the Eisenberg solvation free energy derivatives into the force field in the hopes of improving predictions for surface residues (C. Schiffer, personal communication).

The combinatorial aspect of the Ponder and Richards work has been employed by Wilson *et al.* (Wilson *et al.*, in press) for predicting relative free energies of binding of peptide substrates to enzymes. This method, which can also be applied to the side chain prediction (Wilson *et al.*, in preparation) problem, uses the Ponder and Richards rotamer library to search conformation space. The side chain rotamers of a cluster of amino acids are substituted at once. All possible rotamer combinations at that cluster are then evaluated using the AMBER force field and an implementation of the Eisenberg-McLachlan solvation free energies. The last section of this chapter describes the this method in greater detail.

Part B. Modeling the structure of the cercarial elastase from *Schistosoma mansoni*

Introduction

Schistosomiasis affects one quarter of a billion people. The disease is concentrated primarily in the tropical areas of northeastern Africa, China, and Brazil. It is estimated that 60% of the inhabitants of the Nile Valley are infected with the blood fluke, *Schistosoma*, which causes the disease. The life cycle of the parasite is quite complex including stages in both snail and mammalian hosts. Schistosomal infection occurs when humans bathe or wade in infested water. Organisms in the cercarial (larval) stage of life follow a temperature gradient to the wading human and then invade the circulatory system by burrowing through the skin. Once in the circulatory system, the schistosomes mature, reproduce, and lay many eggs. The eggs can become lodged in many of the human host's tissues. As a defense mechanism, the body forms cysts around the eggs, creating scar tissue in previously healthy organs. The eggs are also excreted in the feces and urine. If they then end up in water inhabited by snails, they invade the snail and eventually develop into the free-swimming cercarial form once again.

Dr. James McKerrow's group at UCSF has purified and cloned a protease (McKerrow *et al.*, 1985) which is expressed during the cercarial stage of the schistosome's life. This protease, termed an elastase because of its ability to cleave elastin, is implicated in the invasion of the human host. It is postulated that the parasite uses this enzyme, along with perhaps others, to chew its way through the extracellular matrix of the dermis. The cercarial elastase has been shown *in vitro* to cleave keratin, laminin, fibronectin, and type IV collagen as well as elastin (McKerrow *et al.*, 1985).

In collaboration with Dr. McKerrow and co-workers, we undertook a project to build a three-dimensional model of the cercarial elastase. It was our hope that visualization of the active site of the molecule might lead to the design of inhibitors of the enzyme. Although inhibitors of this enzyme would act to prevent infection rather than to cure or treat victims of the disease, the design of a compound which could suppress

skin invasion could lead to the development of a topical ointment, similar to a sun-screen or insect repellent. Currently, a lotion preparation of the compound niclosamide is being tested by the U.S. Army for efficacy against cercarial invasion (Cherfas, 1989).

Studies with a battery of mechanism-based inhibitors suggested that cercarial elastase was likely to be a serine protease (McKerrow *et al.*, 1985). The sequence of cercarial elastase was found to be similar to the trypsin-like class of serine proteases (McKerrow *et al.*, 1985). Sequence identity was quite high in the region of the three residues of the catalytic triad (His-57, Asp-102, and Ser-195 -- chymotrypsin numbering scheme; His-41, Asp-99, Ser-191 -- cercarial elastase sequential numbering), but rather low overall -- in the neighborhood of 20%. Cercarial elastase was similar in sequence identity to both the eukaryotic serine proteases and the bacterial serine proteases. Figure 4.1 shows an evolutionary tree relating different proteases. This tree places cercarial elastase as being distantly related to all proteases.

Methods

Modeling the structure of cercarial elastase

We chose to use the available mammalian serine proteases in the Brookhaven Protein Data Bank (PDB) (Abola *et al.*, 1987; Bernstein *et al.*, 1977) as a basis set for modeling the cercarial enzyme. Although we could have chosen to use the bacterial proteases as a basis set, cercarial elastase is equally similar in sequence identity to both groups and is closer in size to the mammalian proteases (having 237 amino acids). At the time of modeling, the structures of six different mammalian or mammalian-like† proteases had been deposited to the data bank. These were bovine trypsin (3PTN), porcine pancreatic elastase (3EST), bovine chymotrypsin (4CHA), rat mast cell protease (3RP2), porcine kallikrein (2PKA) and *S. griseus* trypsin

† mammalian-like refers to *S. griseus* trypsin, which is more closely related in sequence and structure to the mammalian proteases than to the other bacterial proteases in the data bank.

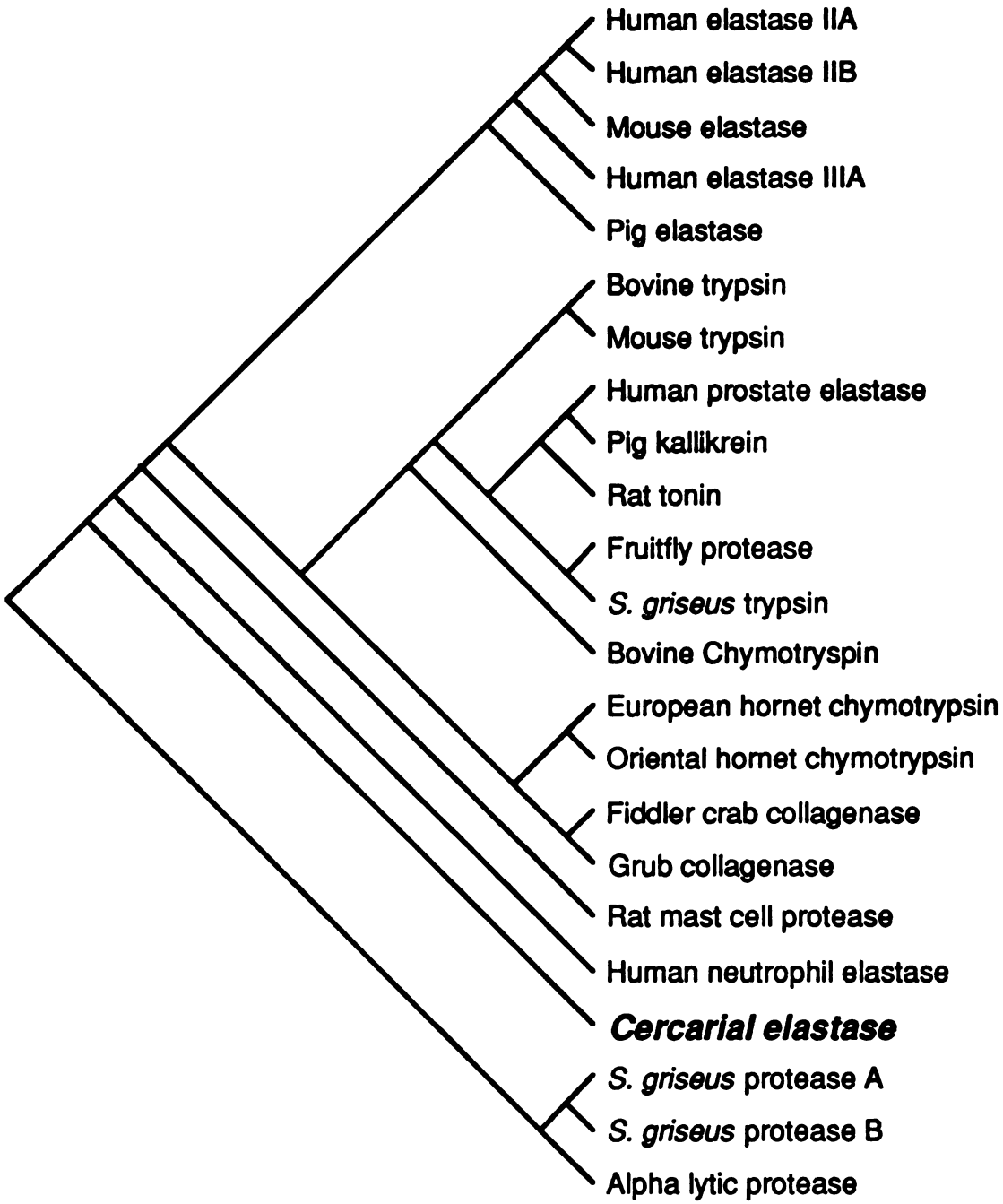


Figure 4.1. Evolutionary relationships between serine proteases. This tree was constructed using the method of parsimony. Sequences were first aligned using the method of Smith and Smith (1990). The tree was generated using the Macintosh program PAUP by Dave Swofford of the Center for Biodiversity, Illinois Natural History Survey. The method of parsimony constructs trees of minimum length by determining the minimum number of nucleotide changes required for one protein to evolve into another. The most parsimonious tree could not be found through an exhaustive search because the large number of sequences made the computation time required too great. The tree was therefore generated using a heuristic method. Three bacterial proteases, alpha-lytic protease, S.griseus protease A and S. griseus protease B, were designated as an outgroup in order to root the tree.

(1SGT). The structure of rat tonin (1TON) was also considered during modeling once its coordinates became available.

The backbones of the structures were superimposed and refined manually with the assistance of computer graphics (UCSF MidasPlus: Ferrin *et al.*, 1988; Jarvis *et al.*, 1988), and from this three-dimensional alignment a sequence alignment was derived. The sequence of cercarial elastase was aligned manually to the multiple alignment of the basis set structures. In most regions, the alignment was fairly straightforward. However, in the region between the catalytic histidine and aspartate, the sequence identity is very low. In general, this region is not very well conserved among the serine proteases. In cercarial elastase, alignment is complicated by a fairly long insertion of approximately 10 or 15 residues. Two different alignments appeared equally reasonable to the eye. One placed the insertion closer to the catalytic histidine and the other closer to the catalytic aspartate. At first we tested the alignments by generating preliminary models. Using computer graphics and the backbone of porcine elastase as a model, we made the appropriate substitutions of amino acids (using the command "swapaa" in UCSF Midas) to correspond to the two alternative alignments. We tried to judge the alternatives on the basis of burial of hydrophobic residues, but both alignments seemed equally plausible. To help us decide where to place the insertion, we used the programs Lineup, Profile, and ProfileGap from the University of Wisconsin UWGCG package to generate an alignment automatically. The program ProfileGap placed the insertion closer to the catalytic aspartate. This alignment was one of the two manually-generated alternatives. The final alignment is shown in Figure 4.2. Table IV.1 shows the sequence identity between cercarial elastase and the other proteases used in modeling.

The model of cercarial elastase was built using the backbone of porcine pancreatic elastase. Elastase was chosen because it was the closest in length (240 residues) to cercarial elastase. Side chains were substituted in their statistically most frequently observed conformations (Ponder & Richards, 1987), except in the region of the sub-

strate binding site. Here, side chains were modeled to match the conformations seen in the basis set structures. In earlier, more preliminary models we made a greater effort to model all side chains in conformations which matched those of the known structures. However, we found that this level of detail was not necessary for the subsequent inhibitor modeling studies. We also subjected our earliest model to AMBER energy minimization (Singh *et al.*, 1987; Weiner *et al.*, 1984) primarily in order to improve the lengths of disulfide bonds. Energy minimization introduced additional errors into our model by excessively distorting the backbone. Therefore, we did not minimize the ultimate model. Substrate binding was modeled using the coordinates of alpha-lytic protease with the boronic acid inhibitor Ala-Ala-Pro-Phe-BOH (Bone *et al.*, 1989) (PDB entry 1P08). The inhibitor was positioned in the binding cleft of cercarial elastase by superimposing the atoms of the catalytic residues in alpha-lytic protease and cercarial elastase.

Our collaborators performed a number of experiments designed to test both the model and itself and predictions regarding the substrate specificity of the enzyme. The enzyme was cloned Dr. Johnny Railey and expressed in *E. coli*. Mutants were made using oligonucleotide-based site-directed mutagenesis. Activity and inhibition assays on the native enzyme were performed by Payman Amiri. Both standard enzyme kinetics and skin penetration assays were performed. Conditions for assays are described in Cohen *et al.* (submitted).

```
          1                                                  50
CercElast IRSGEPVQHP AEFPFIAFLT TERTMCTGSL VSTRAVLTAG HCVCSPLPVI
CONSENSUS ivgGtea..p nswP.qvslq ...hfCGGsL inq.wVltAA HC.....itv
  Trypsin IVGGYTCG.A NTVPYQVSLN S.YHFCGGSL INSQWVVSAA HC.....IQV
SGTrypsin VVGGTRAA.Q GEFPFMVRLS M...GCGGAL YAQDIVLTAA HC.....ITA
  RMCprot IIGGVESI.P HSRPYMAHLD I...ICGGFL ISRQFVLTAA HC.....ITV
Kallikrie IIGGRECE.K NSHPWQVAIY H.SFQCGGVL VNPKWVLTAA HC.....YEV
 Elastase VVGGTEAQ.R NSWPSQISLQ Y.AHTCGGTL IRQNWVMTAA HC.....FRV
Chymotryp IVNGEEAV.P GSWPWQVSLQ D.FHFCGGSL INENWVVTAA HC.....DVV

          51                                                 100
CercElast RVSFLTLRNG DQQGIHHQPS GVKVAPGYMP SCMSARQRRP IAQTLSGFDI
CONSENSUS vlGeh.lnq. egt.qk..vt kv.vhp.yN. .......... ........Di
  Trypsin RLGEDNINVV EGNEQFISAS KSIVHPSYNN .......... ......NNDI
SGTrypsin TGGVVDL.QS G.SAVKVRST KVLQAPGYNG .......... ......GKDW
  RMCprot ILGAHDVRKA ESTQQKIKVE KQIIHESYNL .......... ......LHDI
Kallikrie WLGRHNLFEN ENTAQFFGVT ADFPHPGFNS .......... ......SHDL
 Elastase VVGEHNLNQN NGTEQYVGVQ KIVVHPYWNG .......... ......GYDI
Chymotryp VAGEFDQGSS SEKIQKLKIA KVFKNSKYNN .......... ......NNDI

          101                                                150
CercElast AIVMLAQMVN LQSGIRVISL PQPSDIPPPG TGVFIVGYGR DDNDRDPSRK
CONSENSUS mLlkla..as l.sav.v..l p.....aa.g ttcv..GWGl tr.......s
  Trypsin MLIKLKSAAS LNSRVASISL PT....ASAG TQCLISGWGN TKS......S
SGTrypsin ALIKKAQPIN ...SQPTLKI A.....AYNQ TFTVVAGWGA NRE......S
  RMCprot MLLKLEKKVE LTPAVNVVPL PSPSDFIHPG AMCWAAGWGK TGV......P
Kallikrie MLLRLQSPAK ITDAVKVLEL PT....PELG STCEASGWGS IEP......E
 Elastase ALLRLAQSVT LNSYVQLGVL PRAGIILANN SPCYITGWGL TRT......Q
Chymotryp TLLKLSTAAS FSQTVSAVCL PSASDDFAAG TTCVTTGWGL TRY......N

          151                                                200
CercElast NGGILKKGRA TIMECRHATN GNPICVKAGQ NFGQLPAPGD SGGPLLPSLQ
CONSENSUS tpdtlq.a.l p.ls..aCk. .sm.CaGy.. ......c.GD SGGPlvck..
  Trypsin YPDVLKCLKA PILSDSSCK. .NMFCAGY.. ......CQGD SGGPVVCS..
SGTrypsin QQRYLLKANV PFVSDAACR. .EEICAGY.. ......CQGD SGGPMFRK..
  RMCprot TSYTLREVEL RIMDEKACV. .FQVCVGS.. ......FMGD SGGPLLCA..
Kallikrie FPDEIQCVQL TLLQNTFCA. .SMLCAGY.. ......CMGD SGGPLICN..
 Elastase LAQTLQQAYL PTVDYAICS. .SMVCAGG.. ......CQGD SGGPLHCL..
Chymotryp TPDRLQQASL PLLSNTNCK. .AMICAGA.. ......CMGD SGGPLVCK..

          201                                        240
CercElast GPVLGVVSHG VTLPNLPDII VEYASVARML DFVRSNI
CONSENSUS g.l.GivSwg s.gca.Pgv. ..ytrvs.yv swinqtiasn
  Trypsin GKLQGIVSWG S.GCA.PGV. ..YTKVCNYV SWIKQTIASN
SGTrypsin WIQVGIVSWG Y.GCA.PGV. ..YTEVSTFA SAIASAARTL
  RMCprot GVAHGIVSYG HPDAK.PAI. ..FTRVSTYV PWINAVVN..
Kallikrie GMWQGITSWG HTPCG.PSI. ..YTKLIFYL DWIDDTITEN
 Elastase YAVHGVTSFV S.GCN.PTV. ..FTRVSAYI SWINNVIASN
Chymotryp WTLVGIVSWG SSTCS.PGV. ..YARVTALV NWVQQTLAAN
```

Figure 4.2 Alignment of cercarial elastase with six proteases of known structure. Only structurally-superimposable residues are shown from the proteins of known structure. Abbreviations: CercElast: cercarial elastase; CONSENSUS: consensus sequence generated using the program LineUp from the UWGCG package; Trypsin: bovine pancreatic trypsin; SGTrypsin: trypsin from *S. griseus*; RMCProt: mast cell protease from rat; Kallikrei: porcine kallikrein; Elastase: porcine pancreatic elastase; Chymotryp: bovine pancreatic chymotrypsin.

Table IV.1 Amino acid identity matrix for structural alignment of serine proteases.

	3EST	3PTN	4CHA	2PKA	3RP2	1SGT	CERC
3EST	-	36.7	38.9	31.8	31.5	30.2	19.7
3PTN	85	-	42.6	37.4	32.2	30.5	20.9
4CHA	91	96	-	32.2	30.1	29.7	19.4
2PKA	75	85	74	-	32.0	25.1	17.5
3RP2	73	72	68	73	-	24.2	24.3
1SGT	70	68	67	57	54	-	20.9
CERC	47	48	45	41	56	48	-

The upper half of the matrix shows the percentage of identical residues in identical positions. The lower half of the matrix is the raw number of identities. Values for cercarial elastase (CERC) are based on a sequence alignment. (Abbreviations: 3EST porcine elastase, 3PTN bovine trypsin, 4CHA bovine alpha-Chymotrypsin, 2PKA porcine kallikrein A, 3RP2 rat mast cell protease II, 1SGT S. griseus trypsin)

Results

Mutagenesis

It is currently difficult to express large quantities of cercarial elastase in *E. coli*. It appears that the enzyme is toxic to the host cell. Different expression strategies are currently being pursued, including the use of a eukaryotic host (yeast) (J. McKerrow, personal communication). Consequently, the proposed experiments listed below have yet to be completed.

1) Cercarial elastase shows a preference for large hydrophobic amino acids at the P_1 site (McKerrow *et al.*, 1985). Pancreatic elastase, however, prefers alanine as a substrate. A comparison of the aligned sequences shows that all proteases have glycine at position 216 (210 cercarial elastase sequential numbering), except porcine pancreatic elastase, which has valine. To see if the substrate specificity of cercarial elastase could be altered to be more pancreatic elastase-like, residue Gly 210 to was mutated to valine.

2) A very early model suggested that Asp-218 was at the bottom of the P_1 binding pocket. We proposed to mutate this residue to both lysine and glutamate to see if binding affinity would be affected. Later, the automatic alignment generated using the Wisconsin package placed this residue with position 223 in the chymotrypsin numbering scheme. This is a surface position, much more likely for a charged residue.

3) In order to help us position the large insertion between the catalytic histidine and aspartate of cercarial elastase, a deletion mutant was made in one of the two regions postulated to be the inserted loop. Residue 52 of cercarial elastase was deleted. In the current model this residue is predicted to be in a β-strand. At the time, we thought that the deletion of a loop residue would not disrupt the structure, while deletion of a residue internal to a beta-strand would significantly lower the melting temperature of the enzyme perhaps to such a degree that the enzyme would not be stable. Recent work by Shortle and Sondek (1990) has shown that insertions can be accommodated in many places in Staphylococcal nuclease. This suggests that deletions may be easier to accommodate than previously thought.

Substrate Specificity

Various predictions were made based on the three-dimensional model of cercarial elastase about the substrate specificity of the enzyme. By examining the model and comparing it to chymotrypsin, we made various suggestions as to which series of inhibitors and substrates to assay. These experiments are described in detail in Cohen *et al.*, 1991. The kinetic assays and skin penetration assays were performed by Payman Amiri.

The P_1 site: It was determined fairly early that cercarial elastase showed a preference for large hydrophobic side chains such as phenylalanine and leucine (McKerrow *et al.*, 1985). Figure 4.3 shows phenylalanine in the P_1 pocket. Our modeling predicted that a side chain as large as tryptophan's could fit in this site, with a slight energetically unfavorable rotation in χ_2. A series of substrates and inhibitors of increasing size were studied. Table IV.2 shows the results of the inhibition assays and Table IV.3 shows the results of the substrate hydrolysis assays.

Table IV.2 Kinetic constants for inhibitors of cercarial elastase

Inhibitor ($-P_4P_3P_2P_1-$)	K_i (μM)	k_3 (s^{-1} x 10^3)	k_3/K_i ($M^{-1}s^{-1}$)	% inhibition of skin invasion
Suc-AAPA-CMK	no inhibition	-	13*	0
Suc-AAPV-CMK	no inhibition	-	0.7*	not done
Suc-AAPL-CMK	12	18	1485	80
Suc-AAPF-CMK	13	11	798	80
Suc-AAPW-CMK	20	10	493	not done
Suc-AKPF-CMK	7	37	563	53
Suc-WAPL-CMK	2	8	3846	not done
Suc-FAPF-CMK	1	6	5483	80
Suc-WAPF-CMK	12	6	521	not done

Abbreviations used: Suc = mehtoxysuccinyl blocking group; CMK = chloromethyl ketone; K_i: inhibition constant; k_3: effective rate constant of inhibition.
* These values are $k_{observed}/\{I\}$ which are equal to k_3/K_i since $\{I\} << K_i$.

Table IV.3 Kinetic constants for substrates of cercarial elastase

Substrate ($-P_4P_3P_2P_1-$)	K_m (μM)	k_{cat} (s^{-1})	k_{cat}/K_m ($M^{-1}s^{-1}$)
Suc-AAPA-Sbzl	no activity	-	-
Suc-AAPV-Sbzl	very low activity	-	-
Suc-AAPL-Sbzl	464	7.5	16200
Suc-AAPF-Sbzl	96	19.4	202100
Suc-AAPW-Sbzl	20	10	493
Suc-AAPV-pNA	very low activity		
Suc-AAPI-pNA	very low activity		
Suc-AAPnL-pNA	300	0.02	56
Suc-AAPL-pNA	118	0.33	2800
Suc-AAPM-pNA	300	0.05	185
Suc-AAPF-pNA	119	0.19	1600

Abbreviations: Sbzl = benzyl thioester; pNA = p-nitroanilide

As expected, large, hydrophobic amino acids are preferred at the P_1 site. For substrate hydrolysis, the optimal side chain size appears to be phenylalanine. Leucine at P_1 makes the best chloromethyl ketone inhibitor. Tryptophan defines the size limit of the P_1 pocket: the substrate analog Suc-AAPW-Sbzl is a poor substrate, but the chloromethyl ketone inhibitor Suc-AAPW-CMK is still reasonable, with a 20 micromolar inhibition constant.

As noted previously (McKerrow *et al.*, 1985), beta-branching of the P_1 amino acid significantly reduces activity. According to the model, Pro-188 may be responsible for specificity against beta-branched substrates. This residue occurs in a loop extension unique to the cercarial proteases. Other proteases have a cysteine at position 187 (cercarial numbering) which could pull the loop away from the binding pocket. The cercarial enzyme has alanine at this position.

The P_2 site: Proline is in the P_2 site. This residue was not altered in the test compounds since it was thought that the restricted geometry enhanced binding. For alpha-lytic protease, substituting alanine for proline at this position reduces k_{cat}/K_m by a factor of ten (Bone *et al.*, 1987).

The P_3 site: The P3 site is solvent exposed. We predicted that a hydrophilic residue at this position should improve solubility, while not affecting binding affinity, so an inhibitor with lysine at this position (Suc-AKPF-CMK) was tested. This inhibitor had a lower K_i than the "parent" AAPF inhibitor and was only slightly less effective at inhibiting the protease (k^3/K_i = 563 $M^{-1}s^{-1}$ versus 798 $M^{-1}s^{-1}$). It was also 4 times more soluble in water and at this higher concentration, inhibited more cercariae from penetrating skin.

The P_4 site: As compared to chymotrypsin, the P_4 site is much more exposed in cercarial elastase -- a loop which hangs over this site is missing in our enzyme. We speculated that a large side chain on the inhibitor or substrate could fit at the P_4 subsite (see Figure 4.4). To test this prediction, substrates with large, hydrophobic amino acids (Phe, Trp) at this position were first assayed. These large amino acids

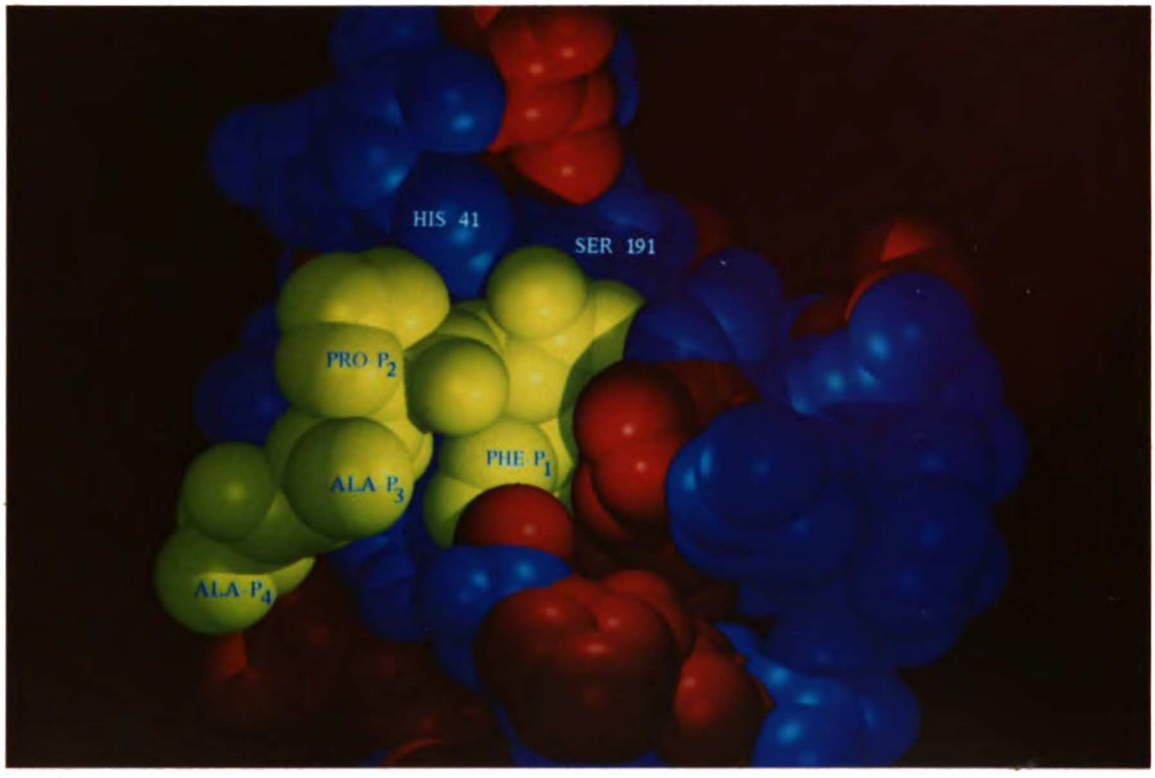


Figure 4.3. Space filling model of active site of the cercarial proteases with substrate Ala-Ala-Pro-Phe. Hydrophobic amino acids are colored red and hydrophilic amino acids are colored blue. Catalytic residues Ser-191 and His-41 are indicated.

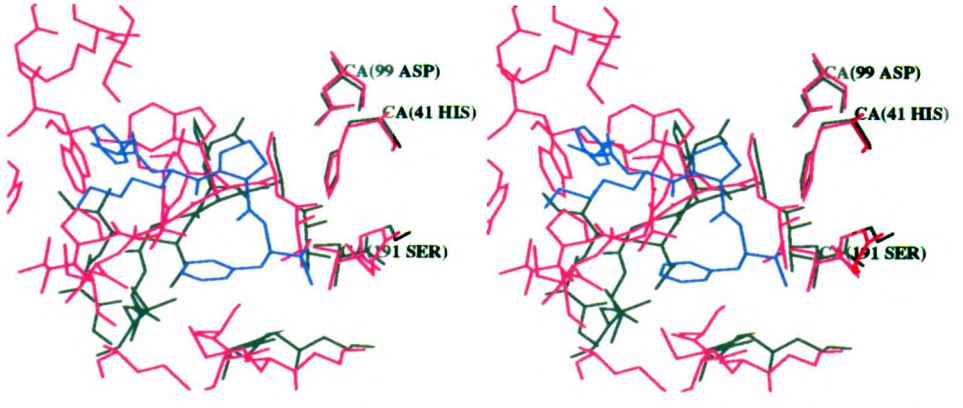


Figure 4.4. Comparison of cercarial elastase (green) and chymotrypsin (magenta) with substrate Phe-Lys-Pro-Phe (cyan). The P_4 site in chymotrypsin is crowded with bulky tryptophan residues while cercarial elastase lacks the large loop which hangs over the S4 site. Lysine is shown at P3 -- a large hydrophilic residue is easily accommodated at this position.

substantially abolished or diminished substrate hydrolysis (Table IV.3). However, chloromtehyl ketone inhibitors with tryptophan at P_4 and phenylalanine or leucine at P_1 were worked well (Table IV.2). We speculate that the interaction of large hydrophobic residues at P4 with the residues lining the P_4 pocket distorts the geometry of the scisslie bond relative to the active site without destroying binding affinity for the inhibitor.

Discussion

Modeling the structure of the skin-penetrating enzyme from *S. mansoni* is an example of the modeling and subsequent testing of a predicted protein structure. The enzyme activity assays, in particular, are instructive of what one can do with a model-built structure. Predictions which may otherwise not have been considered regarding substrate specificity can be made and tested. For example, we proposed that a long, hydrophilic side chain could work at the P_3 site. Indeed, a lysine at this position in an inhibitor does not diminish binding affinity significantly, but dramatically increases aqueous solubility. Similarly, we noted that our model lacks a large loop which constricts the P_4 site and predicted that large amino acid side chains might be tolerated here. Although substrates with large side chains at P_4 were not very good in terms of k_{cat}/K_m, inhibitors bound with high affinity. Since our ultimate goal is to inhibit this enzyme to prevent schistosomal infection, this model has provided interesting leads to follow in the development of an inhibitor.

Part C. Improving Side Chain Positioning

Introduction

In 1987 Jay Ponder and Fred Richards published a paper describing a rotamer library approach for determining "tertiary templates" for protein structures (Ponder & Richards, 1987). Ponder and Richards had hypothesized that an important determinant for folding was satisfactory packing of the core of the protein. They postulated that they could predict which sequences were compatible with a particular tertiary fold by determining which residues could pack in the core. Both volume effects and steric effects were considered. A tertiary template is therefore a list of permitted sequences of core residues. To determine the tertiary template for a core structure, a packing unit of approximately five residues is chosen. All low energy conformations, or rotamers, for all amino acids are combinatorially substituted. If the side chains overlap excessively, or if large cavities are found, the rotamer combination is discarded. Suitable residue sets for the packing site are those which fill the volume efficiently without steric clashes.

Although not its original intent, the work of Ponder and Richards suggests a natural method for predicting the conformations of side chains. Described here is a homology modeling tool which builds side chains onto a structure and optimizes their conformations. This method, orignially applied to the prediction of relative binding free energies of peptide inhibitors to mutant alpha-lytic proteases (Wilson *et al.*, in press), uses the Ponder and Richards rotamer library of side chain conformations. The idea of packing sites is also implemented: unlike other side chain conformation prediction methods (Novotny *et al.*, 1988; Schiffer *et al.*, 1990), this method is multidimensional and optimizes the conformations of of a group of nearby residues at once.

This work was a collaborative project with Charles Wilson. He developed the side chain conformation prediction algorithm, while I built the test structures by homology modeling and developed the error analysis methodology. Parts of the Methods,

Results, and Discussion sections are taken from Wilson *et al.*, in preparation, which we wrote together.

Methods

Side chain prediction algorithm

The first step in the side chain prediction algorithm is the calculation of coordinates for all rotamers for all of the amino acids in the structure. The coordinates are computed using the Ponder and Richards program, PROPAK and the published rotamer library is used. Rotamers which make bad contacts with main chain atoms are pruned. After generating the rotamers, the following step-by-step procedure is applied† :

1) One of the amino acids in the model is chosen at random as a site center.
2) The five residues whose side chains are the closest to the site center are identified.
3) For the six residues in the site, all possible rotamer combinations are tested, and for each combination, and 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. Convergence does not always occur because sites overlap and a given side chain may have different optimal conformations in several sites. Therefore, the procedure is stopped after having cycled through the protein three times.

The force field used to evaluate rotamer combinations contains two parts: 1) non-bonded interactions between atom pairs and 2) the change in solvation energy upon exposing the atoms to solvent. The non-bonded terms (including electrostatic, van der

† The following computations are all done on the MSG VAX using the program REC by Charles Wilson (see Wilson *et al.*, in press)

Waals and hydrogen bond energies) are calculated using the AMBER force field (Weiner *et al.*, 1984) with a distance-dependent dielectric constant, $\varepsilon = r$, in the electrostatic term.

The solvation energy is calculated using a model similar to that of Eisenberg and McLachlan (Eisenberg & McLachlan, 1986) (see Wilson *et al.*, in press). In their formalism, each atom is assigned an atomic solvation parameter (ASP) and the solvation energy is the product of an atoms accessible surface area and its ASP. Instead of computing the solvent accessible surface area for each atom (a time-intensive computation), a grid method is used to estimate solvent accessibility. Solvent molecules are represented as grid points on a 1.0 Å body-centered cubic lattice which surrounds the protein. The number of grid points surrounding an atom is proportional to the total atomic accessible surface area and this to the solvation energy for the atom.

Construction of initial model structures

To evaluate the side chain modeling procedure, I constructed several "homology-built" models using pairs of known structures. The pairs covered a wide range of sequence similarities (from 30 to 100%). The pairs tested are listed in Table IV.4. The starting models were constructed with the assistance of computer graphics (UCSF MidasPlus: Ferrin *et al.*, 1988; Jarvis *et al.*, 1988). A correct sequence alignment was made by first generating a structural alignment of the "true" and "template" structures. The true structure is the known structure of the protein whose side chain conformations will be modeled, and the template structure is the structure from which the model is going to be built. A correct alignment had to be assumed in order to controllably test the side chain modeling algorithm. In a real-life modeling situation, where one has a known three-dimensional structure and a sequence of interest, the *sequences* must be aligned. If the sequence identity is less than 70%, the alignment, although evolutionarily reasonable, may be structurally incorrect.

In the MacroMolecular WorkBench group at UCSF, we currently have no automated method of generating three-dimensional structural alignments. In this study,

the structures were superimposed "manually". The two structures were first aligned structurally on the graphics screen by eye. Refinement of the structural alignment was made using successive calls of the "match" command in MidasPlus, which performs a least squares fit of a minimum of four atoms from each structure. When the alignment of the backbones appeared optimal, the coordinates of the aligned backbones were saved. The program STRUCTALIGN was written to generate a sequence alignment from the overlapping coordinate sets. STRUCTALIGN generates a difference distance matrix for all alpha-carbon - alpha-carbon pairs from the two structures. For each

Table IV.4: Test cases for side chain conformation optimization

model (template→ unknown)	N_{unk}	N_{tmpl}	N_{mod}	overall identity (%)	identity of modeled regions (%)	resol. unk (Å)	resol. tmpl (Å)	backbone r.m.s.d. (Å)
ALP→ALP	198	198	198	100.0	100.0	1.7	1.7	0.00
LBP→LIV	344	346	344	79.1	79.4	2.4	2.4	0.69
LZ1→LYZ	129	130	129	60.2	60.5	1.5	2.0	0.61
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 (1990); **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).

alpha-carbon in the first structure, the closest alpha-carbon in the second structure within 2.5 Å is located. The sequence alignment is printed. If no alpha-carbons are within 2.5 Å, no partner is assigned for that particular residue. The output of STRUCT-ALIGN was modified by hand. Occasionally, two residues which are clearly homologues will be missed because they are in similar but shifted loops, and thus outside the 2.5 Å cutoff. Generally, spatially distant loops, or loops with large insertions or deletions were not included in the models. In other instances, two sequential residues from the first structure will be assigned to the same residue in the second structure because the alpha-carbon in the second structure is the closest atom to both alpha-carbons in the first structure.

Given the alignment, the residues of the to-be-modeled structure were substituted for the residues of template structure. All residues were attached in their most frequently observed conformation as tabulated by Ponder and Richards (Ponder & Richards, 1987). Amino acid substitution was performed using the "swapaa" command in MidasPlus. A script of swapaa commands was written in order to automate this process.

Evaluation of model structures

The final predicted models were evaluated on the basis of rms deviation using the backbones of the template and unknown structures for superposition. Both side chain and main chain rms deviations for each residue were computed. The overall side chain rms deviation was computed as well. These calculations were done using the program SCRMS (which I wrote.)

I generated a "best possible" structure using the template structure backbone and the rotamer library (also using SCRMS). At each position, the rotamer having the lowest r.m.s. deviation from the side chain in the true structure was selected. This structure is the best result we could hope to obtain with the algorithm since we allow idealized rotamers for the side chains. To score our predictions, we determined the fraction of side chains assigned to the same rotamer as in the best possible structure.

To determine whether the procedure actually improves the accuracy of the model, we generated structures which had the most common rotamer (as tabulated by Ponder and Richards) assigned to each residue. The r.m.s. deviation of this unrefined model from the true structure gives a baseline measure against which to compare other models. These structures are referred to as the "first guess" structures, since placing side chains in their most common conformations good first approximation.

Results

An idealized test case:

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 isolate the possible sources of error as much as possible, the following changes in the modeling procedure were made. 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. 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 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 IV.5). While the majority of side chains are modeled 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 surprising 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). For 75% 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 second lowest energy rotamer is the correct choice for 13 of 16 incorrect side chains. The incorrect residues include an unusually high number of serines (6) and asparagines (5), indicating perhaps that uncharged hydrogen bonds may be treated improperly by the force field.

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 residues Asn 62, Asn 118, Asn 162, and Ser 189, the calculated energy difference between the correct side chain rotamer and the lowest energy rotamer is more than 10 Kcal/mole — it seems unlikely that force field errors alone could account for this large difference. 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 in the crystal structure for the four incorrectly-predicted residues with large energy errors. It is possible that these errors

Table IV.5: Results of the α-lytic protease test cases

Conditions of the test	average r.m.s. deviation (Å)	overall r.m.s. deviation (Å)	fraction correct (correct/total)
1. Starting with the crystal structure including symmetry-related molecules and counter-ions; replaced 'best' rotamers with true side chains; varied a single side chain at a time; structure corrected after each rotamer search.	0.26 ± 0.58	0.59	0.89 (126 / 142)
2. As with (1) but lacking symmetry-mates and counter-ions.	0.27 ± 0.61	0.62	0.89 (126 / 142)
3. As with (1) but lacking symmetry-mates and counter-ions; no replacement of 'best' rotamers with true side chains.	0.68 ± 0.85	1.21	0.82 (116 / 142)
4. Starting with no side chains; lacking symmetry-mates and counter-ions; no replacement of 'best' rotamers with true side chains; varyied clusters of five residues simultaneously.	0.73 ± 0.91	1.31	0.76 (111 / 142)

Footnote: 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.

may have been avoided if side chain hydrogen atoms had been included in the x-ray refinement.

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 IV.5). In all subsequent tests, symmetry related atoms and bound sulfates have been ignored.

In the first test case we replaced the best rotamer in the library with 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 rotamer approximation obviously results in additional errors in the model but the effect is small. 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 IV.5) 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.

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 that we have introduced errors in the backbone positions used to place the side chain rotamers. These results are summarized in Table IV.6. 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 4.5).

As with the α-lytic crystal structure test case, the conformations of solvent-accessible residues are signficantly harder to predict than buried residues. Figure 4.6 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 IV.6.

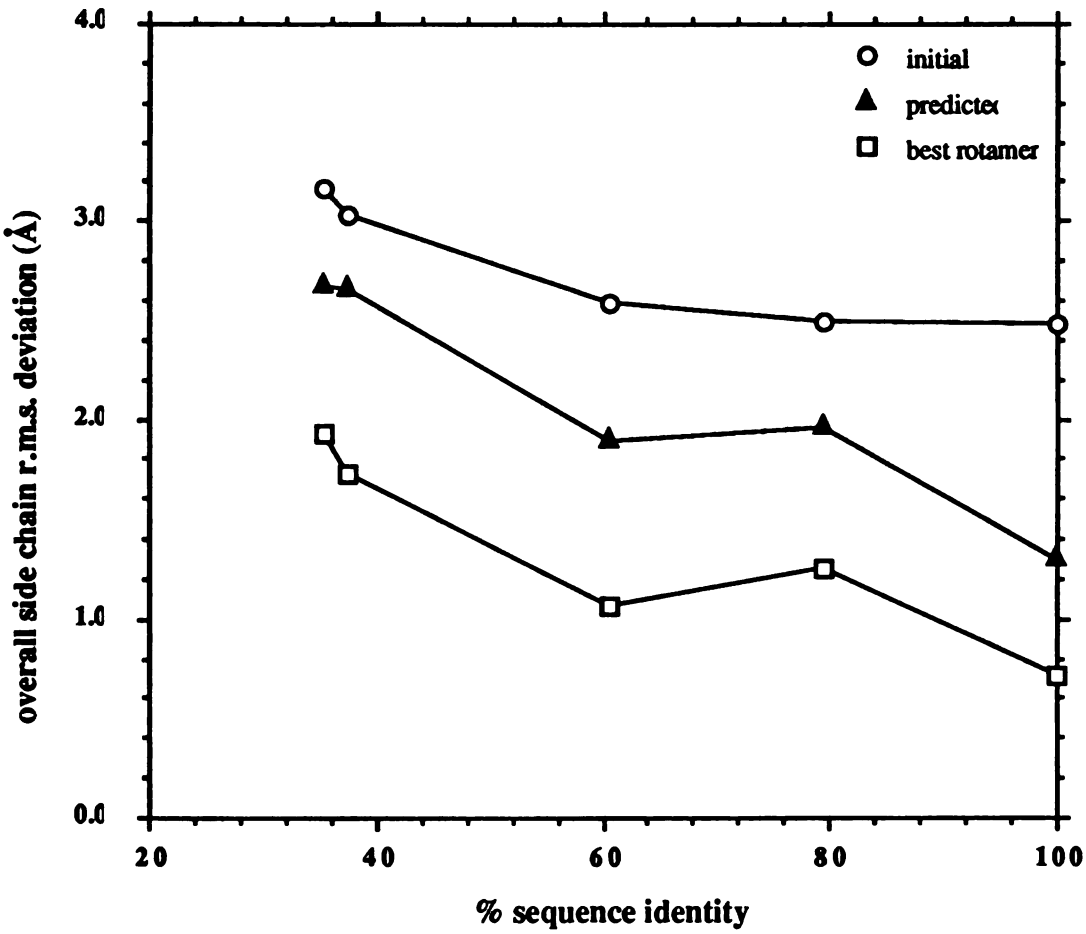


Figure 4.5. Accuracy of the side chain prediction as a function of percent homology. The overall side chain r.m.s. deviation is shown as a function of the percent homology between the unknown and predicted structures for the first guess, predicted, and 'best rotamer' (lowest r.m.s. deviation) models.

Table IV.6: Homology modeling results

structure	first guess average rmsd	first guess overall rms (Å)	predicted average rms (Å)	predicted overall rms (Å)	best average rms (Å)	best overall rms (Å)	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

Footnote: Results for homology modelling test cases. "first guess" structures have the most common rotamer installed at every amino acid position, "predicted" structures have had the combinatorial rotamer search procedure applied, and "best" structures have the rotamers closet in conformation to the true structure installed. Solvent accesibility was calculated using the method of Lee and Richards as implemented in the program ACCESS by Handschumacher and Richards. Residues considered buried have less than 20% of their accessible surface aresase exposed (relative to an extended tripeptide model). Abbreviations for the proteins are the same as in Table IV.5.

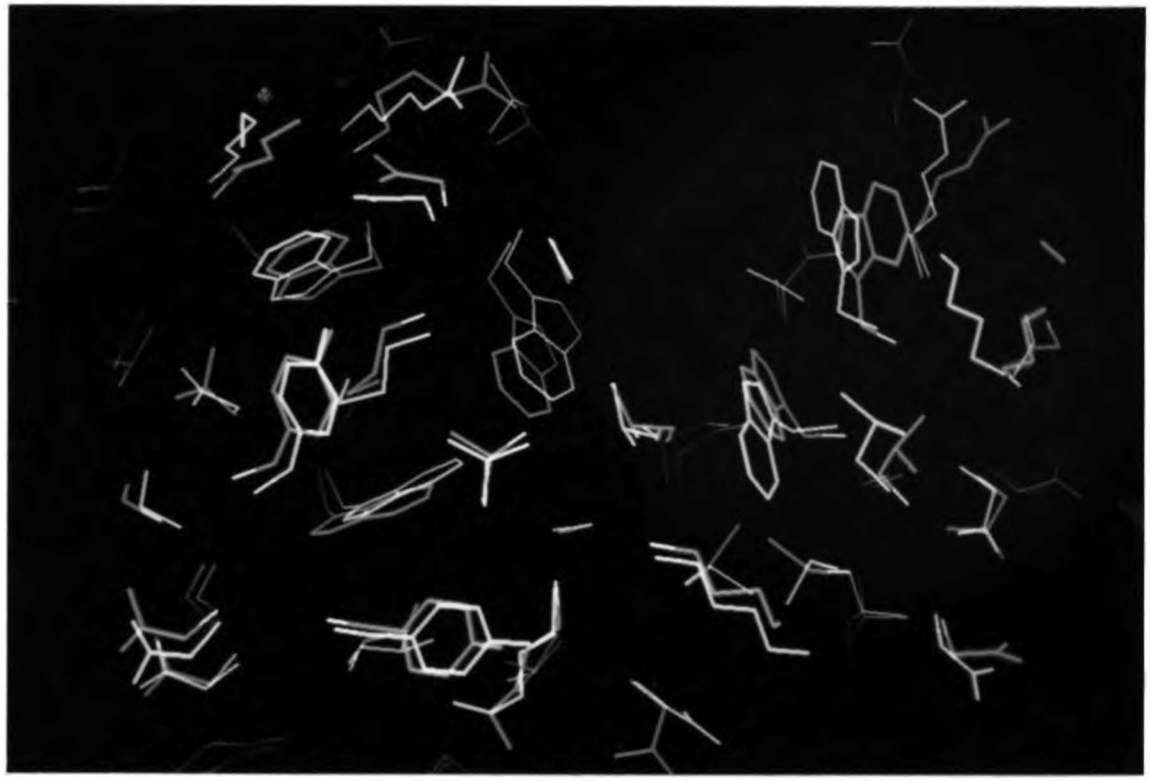


Figure 4.6. Comparison between the predicted and observed hen eggwhite lysozyme structures. Side chain and C_α atoms are shown for the predicted (blue) and true (yellow) structures (C_α atoms colored magenta). Residues in the hydrophobic core lie on the left-hand side while those on the right are generally somewhat solvent accessible.

Increased errors at the surface could be due to additional side chain conformational freedom (since there are fewer restricting adjacent residues), to more crystallographic errors at the surface, or to errors in the force field that effect electrostatic interactions more than van der Waals interactions (since hydrophilic residues are found predominantly at the surface). Previous analysis of protein crystal structures has shown that surface residues have systematically highly 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, the conformations of 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 4.7 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 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). By altering the direction of the $C_\alpha \rightarrow C_\beta$ vectors, however, these shifts cause a significant change in the positions of the calculated rotamers. The r.m.s. deviation of the 'best' 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 best 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 best rotamer structure to the bad contact distance of 2.43 Å in the SGB→ALP structure (figure 4.7). Whereas the 'best' 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.

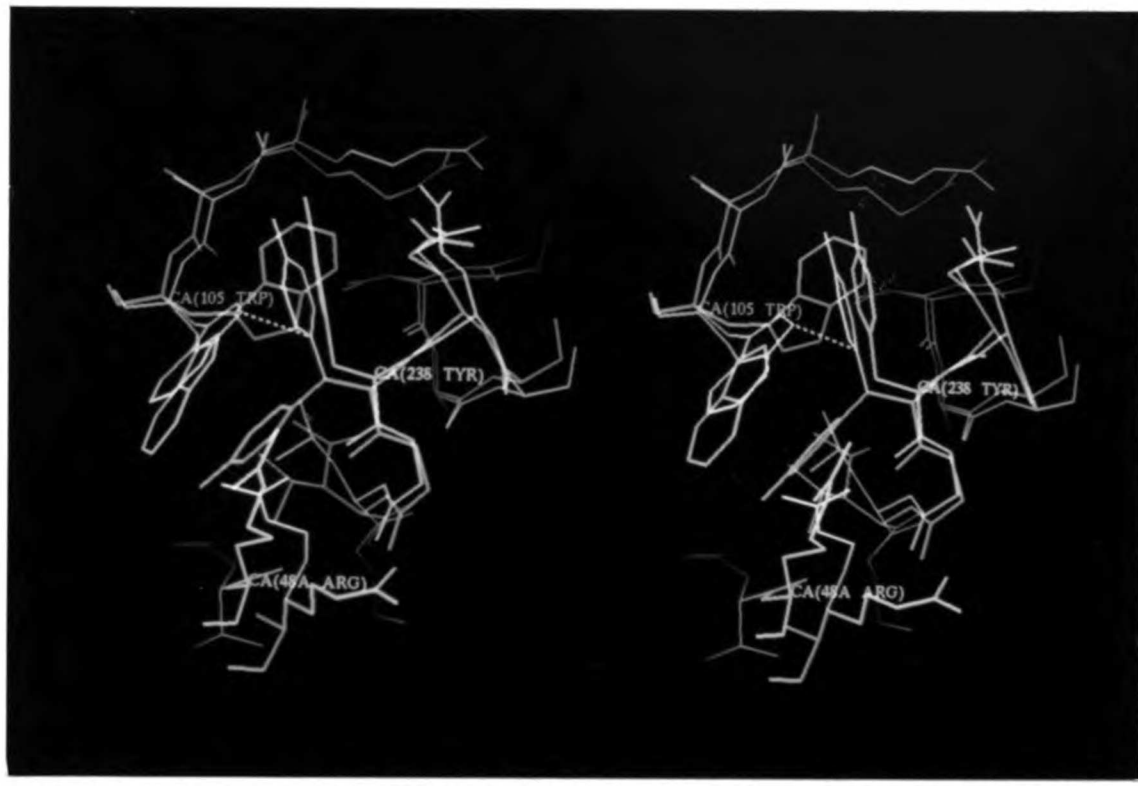


Figure 4.7. Errors in the *S. griseus* protease B → α-lytic protease prediction. The true (yellow) and predicted (blue) structures for α-lytic protease are shown (using *S. griseus* protease B as a backbone template). The 'best' rotamers (those with the lowest r.m.s. deviation to the true structure) 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).

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 modeling test. The fraction of properly placed side chains is a function of the simi-

larity 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 compares favorably, both in terms of accuracy and speed, with energy-based side chain modeling algorithms that have been previously reported (Bruccoleri *et al.*, 1987; Schiffer *et al.*, 1990). 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 loop conformation 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, it is unlikely that the lack of bonded terms presents a problem. 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 conformation space sampling. Whereas the CONGEN program can search an arbitrary number of rotamers for a single side chain, our algorithm combinatorially tests a handful 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). In contrast to the CONGEN method, however, 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 4.7), 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.

References for Chapter 4

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. (1987). In *Crystallographic Databases -- Information Content, Software Systems, Scientific Applications*. (Allen, F.H., Bergeroff, G. & Sievers, R., ed.), pp. 107-132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Alber, T., Sun, D., Nye, J., Muchmore, D. & Matthews, B. (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.

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.

Bone, R., Shenvi, A.B., Kettner, C.A. & Agard, D.A. (1987). Serine Protease Mechanism: Sturcture of an Inhibitory Complex of α-Lytic Protease and a Tightly Bound Peptide Boronic Acid. *Biochemistry*. **26**, 7609-7614.

Bone, R., Silen, J. & Agard, D. (1989). Structural Plasticity Broadens the Specificity of an Engineered Protease. *Nature*. **339**, 191-195.

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

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

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

Cherfas, J. (1989). New Weapon in the War Against Schistosomiasis. *Science*. **246**, 1242-1243.

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

Chothia, C., Lesk, A.M., Levitt, M., Amit, A.G., Mariuzza, R.A., Phillips, S.E.V. & Poljak, R. (1986). The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. *Science*. **233**, 755-758.

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.

Cohen, F.E., Gregoret, L.M., Amiri, P.A., Aldape, K., Railey, J. & McKerrow, J.H. (submitted). Arresting Tissue Invasion of a Parasite with Synthetic Protease Inhibitors Chosen by Computer Modeling. *Biochemistry*.

Dorit, R.L., Schoenbach, L. & Gilbert, W. (1990). How Big is the Universe of Exons. *Science*. **250**, 1377-1382.

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

Ferrin, T.E., Huang, C.C., Jarvis, L.E. & Langridge, R. (1988). The MIDAS Display System. *J. Mol. Graphics*. **6**, 13-37.

Greer, J. (1990). Comparative Modeling Methods: Application to the Family of the Mammalian Serine Proteases. *Proteins: Struct., Func., Genet.* **7**, 317-334.

Jarvis, L., Huang, C., Ferrin, T. & Langridge, R. (1988). UCSF MIDAS: Molecular Interactive Display And Simulation. *J. Mol. Graphics*. **6**, 2-27.

Johnson, M., Sutcliffe, M. & Blundell, T. (1990). Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins. *J. Mol. Evol.* **30**, 43-59.

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

Kneller, D.G., Ph.D. Thesis, Modeling of Loops in Proteins. University of California, Berkeley, 1988.

McGregor, M., Islam, S. & Sternberg, M. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295-310.

McKerrow, J.H., Pino-Heiss, S., Lindquist, R. & Werb, Z. (1985). Purification and characterization of an Elastinolytic Proteinase Secreted by Cercariae of *Schistosoma mansoni*. *J. Biol. Chem.* **260**, 3703-3707.

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

Schiffer, C.A., Caldwell, J.W., Kollman, P. & Stroud, R.M. (1990). Prediction of Homologous Portein Structures Based on Conformational Searches and Energetics. *Proteins: Struct., Func., Genet.* **8**, 30-43.

Shih, H.H.-L., Brady, J. & Karplus, M. (1985). Structures of Protesin with Single-Site Mutations: A Minimum Perturbation Approach. *Proc. Natl. Acad. Sci. USA.* **82**,

Singh, U.C., Weiner, P.K., Caldwell, J. & Kollman, P.A. (1987). AMBER 3.0. University of California, San Francisco.

Smith, R.F. & Smith, T.F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci., USA.* **87**, 118-122.

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., Funct., Genet.* **1**, 267-279.

Sondek, J. & Shortle, D. (1990). Accommodation of single amino acid insertions by the native state of staphylococcal nuclease. *Proteins: Struct., Func., Genet.* **7**, 299-305.

Summers, N.L., Carlson, W.D. & Karplus, M. (1987). Analysis of Side-chain Orientations in Homologous Proteins. *J. Mol. Biol.* **196**, 175-198.

Sutcliffe, M., Hayes, F. & Blundell, T. (1987). Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. *Protein Engineering*. **1**, 385-392.

Taylor, W.R. (1986). Identification of Protein Sequence Homology by Consensus Template Alignment. *J. Mol. Biol.* **188**, 233-258.

Taylor, W.R. & Orenga, C.A. (1989). Protein Structure Alignment. *J. Mol. Biol.* **208**, 1-22.

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.

Wilson, C., Mace, J.E., Agard, D.A. (in press) A Computational Method for the Design of Enzymes with Altered Substrate Specificity, *J. Mol. Biol.*

Wilson, C., Gregoret, L.M. & Agard, D.A. (in preparation). Modeling Side Chain Conformations for Homologous Proteins Using an Energy-Based Rotamer Search.

Chapter 5.

Hydrogen Bonds Involving Sulfur Atoms in Proteins†

† This chapter has been published with co-authors Stephen D. Rader, Robert J. Fletterick, and Fred E. Cohen in the journal Proteins: Structure, Function, and Genetics, volume 9, pages 99-107, 1991.

Introduction

Sulfur is found in proteins in the side chains of the amino acids cysteine and methionine. Cysteine is best known for its unique ability to form cross-links via disulfide bonds. Methionine is usually categorized as an uncommon hydrophobic amino acid. Although a thorough analysis of metal ion binding by these amino acids in proteins has recently been completed (Chakrabati, 1989), little attention has been given to their ability to participate in hydrogen bonding (Baker & Hubbard, 1984) perhaps because of their relative rarity in proteins of determined three-dimensional structure. Gray and Matthews (Gray & Matthews, 1984) have called attention to the role of side chain - backbone hydrogen bonds in helices. We were motivated to survey the frequency and geometry of sulfur-containing hydrogen bonds in globular proteins in order to better assess the importance of this interaction and gauge the interatomic packing interactions of sulfur. Site directed mutagenesis has made it easy to exchange amino acids in a protein. Methods for predicting the effects of various amino acid substitutions on protein structure and function are important for experimental design. A more detailed examination of the interactions particular to specific amino acids should help reach this end.

Reduced sulfur atoms are known on sound theoretical basis to be capable of accepting or donating hydrogen bonds (Kollman et al., 1975). The sulfhydryl group of cysteine can act either as a hydrogen bond donor or as an acceptor. The sulfurs of methionine and half-cystine, lacking hydrogens, can only accept hydrogen bonds. The strength of a hydrogen bond between H_2S and H_2O has been calculated to be 3.1 to 3.2 kcal/mol *in vacuo* when sulfur is the hydrogen bond donor or acceptor (Kollman et al., 1975). In non-covalent enzyme-substrate interactions, the magnitude has been shown experimentally to be slightly smaller: upon replacing a cysteine involved in substrate binding by glycine and serine, Wilkinson and co-workers calculate the decrease in transition state stabilization to be approximately 1.1 kcal/mol (Wilkinson

et al., 1983). The strength of *structural* hydrogen bonds in proteins has not been probed experimentally, but it is presumed to be of similar magnitude.

Hydrogen bonds involving sulfur atoms are longer than those involving nitrogen or oxygen because of sulfur's larger size and more diffuse electron cloud. The equilibrium distance from donor to acceptor atom in a hydrogen bond between a hydroxyl group and an oxygen atom is 2.95 Å, whereas the distance between a sulfhydryl group and an oxygen is 3.66 Å (Kollman et al., 1975). The distance between -SH and O in crystals of L-cysteine is 3.4 Å (Kerr et al., 1975).

Sulfur is instrumental in the active sites of the sulfhydryl proteases such as papain and actinidin (Kamphuis et al., 1985) and in the viral cysteine proteases (Bazan & Fletterick, 1988). These enzymes use Sγ of cysteine as a nucleophile for peptide bond cleavage. McGrath et al. (McGrath et al., 1989) recently substituted serine-195 of rat trypsin with cysteine in order to determine whether trypsin could be engineered to be a sulfhydryl protease.

We have surveyed protein structures for the occurence of hydrogen bonds involving sulfur atoms. The results of this survey underscore the necessity to separate reduced cysteine from disulfide bonded half-cystine in analyzing the three-dimensional coordinates in the protein database. Cysteine behaves differently when it is reduced and when it is part of a disulfide bond. This difference is in part attributable to the differences in hydrogen bonding ability of these two types of cysteines.

Methods

We examined the atomic coordinates of 85 protein structures from the Brookhaven Protein Data Bank (Abola et al., 1987; Bernstein et al., 1977)† . This group was

† Data set of Brookhaven Protein Data Bank structures analyzed:
1ACX, 1ALC, 1BP2, 1CAC, 1CCR, 1CRN, 1CSE, 1ECA, 1FX1, 1GCR, 1GD1, 1GOX, 1GP1, 1HDS, 1HIP, 1HMQ, 1HNE, 1HOE, 1LZ1, 1LZT, 1MB5, 1NXB, 1PAZ, 1PCY, 1PSG, 1RDG, 1RNS, 1SGT, 1SN3, 1TON, 1UBQ, 1UTG, 2ACT, 2ALP, 2APP, 2APR, 2AZA, 2CAB, 2CCY, 2CDV, 2CI2, 2CNA, 2CPP, 2CPV, 2CYP, 2FB4, 2LH1, 2LHB, 2LZM, 2MHB, 2MHR, 2OVO, 2PAB, 2PRK, 2RHE, 2RSP,

selected from structures which were distinct and for which complete atomic detail is available at an experimental x-ray diffraction resolution better than or equal to 2.0 Å. All amino acids containing covalently bonded sulfur atoms (other than half-cystine) and sulfur atoms participating in metal ion binding were not evaluated. Our data set thus consisted of 109 cysteines, 307 methionines, and 268 half-cystines.

All atoms within 4.25Å of each methionine and cysteine sulfur atom were located. This distance was chosen to be long enough to include donor -- acceptor pairs within van der Waals contact of each other, yet short enough to supply meaningful data. For all nearby atoms, excluding those likely to contribute to nonspecific interactions (atoms from the cysteine or methionine in question, the backbone carbonyl carbon and oxygen atoms of the preceding residue, and the backbone nitrogen of the following residue) we calculated the angles and distances defined in Figure 5.1. We then sorted these nearby atoms by atom type or functional group into four categories: carbon, nitrogen, carbonyl oxygen (both backbone and side chain), and hydroxyl oxygen, and prepared distributions of angles and distances for each atom or group. In order to compare the distributions, each was normalized by the number of occurrences of that particular atom or group in the data set and by the number of cysteines, methionines, or half-cystines. The distance distributions were also normalized by shell volume.

Carbon atoms do not participate in hydrogen bonding, yet short distances between non-adjacent carbon and sulfur atoms are observed in the data set of structures. X-ray crystallographic refinement permits some small number of short contacts since the process is a least-squares minimization. Some of these short contacts are real and others are erroneous. The carbon to sulfur distance and angle distributions observed should define background levels for random interactions. In order to look for peculiarities in the distance and angle distributions for actual hydrogen bond donor ($-NH_n$, -

---

2SGA, 2SNS, 2SOD, 2WRP, 3BCL, 3C2C, 3DFR, 3EST, 3FAB, 3GRS, 3INS, 3RNT, 3RP2, 3SGB, 3TLN, 451C, 4FD1, 4FXN, 4HHB, 4PTP, 4RXN, 5CHA, 5CPA, 5CYT, 5PTI, 5TNC, 6LDH, 7RSA, 9PAP

OH) and acceptor (carbonyl O) groups, for each group we prepared a "difference distribution" by subtracting the normalized carbon distance or angle distribution from the distribution in question.

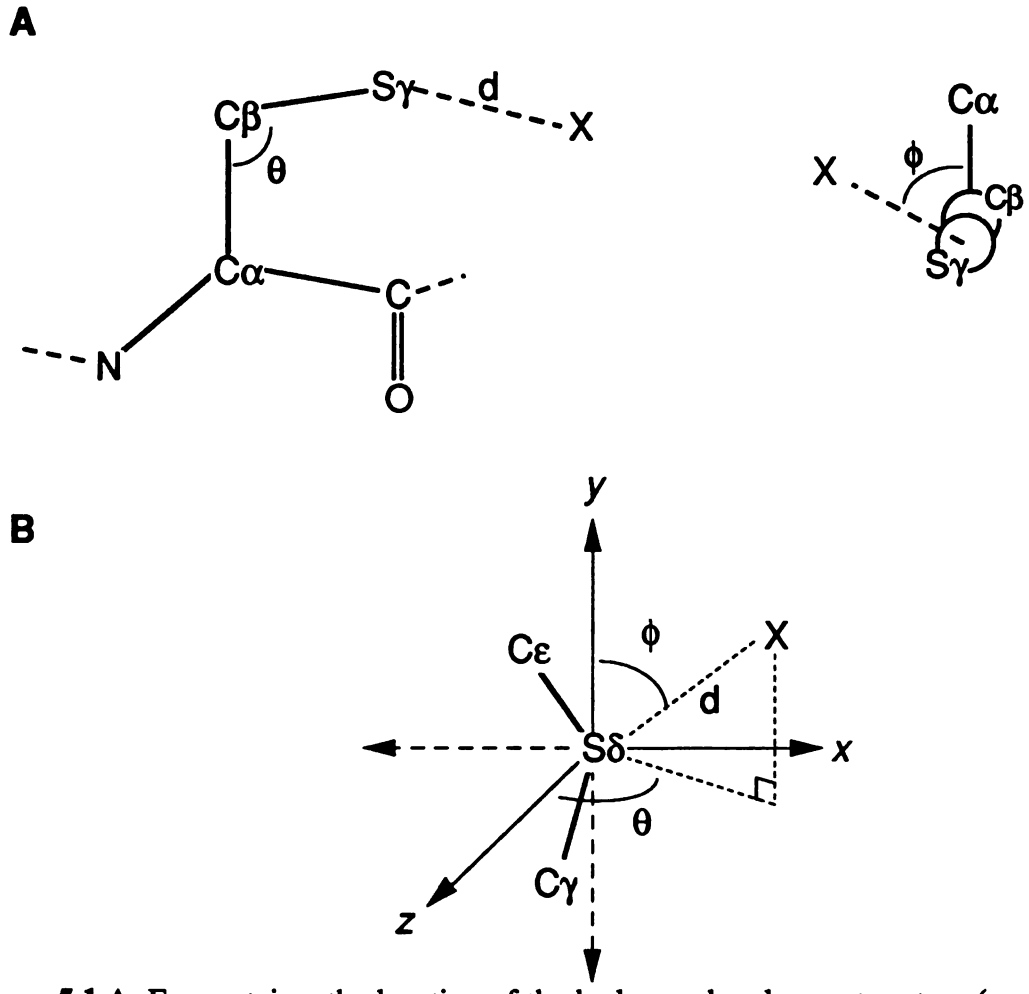


Figure 5.1 A. For cysteine, the location of the hydrogen bond acceptor atom (or donor atom in the case of half-cystine), X, is determined by three parameters: **d**: the distance from the γ-sulfur (Sγ) to the donor/acceptor (X), θ: the angle between the β-carbon of cysteine (Cβ), Sγ and X. and ϕ: the dihedral angle defined by the α-carbon (Cα), Cβ, Sγ, and X. **B.** For methionine, a coordinate system is oriented with respect to the δ-sulfur (Sδ), and the ε and γ carbons (Cε, Cγ). Sδ is placed at the origin and Cε and Cγ are placed in the y-z plane. The bisector of the angle ∠Cγ Sδ Cε is placed on the z-axis. The three parameters defining the position of the donor atom are **d**: the distance from Sδ to X, θ: the angle when X is projected onto the x-z plane, and ϕ: the angle between the y-axis, Sδ and X.

Results and Discussion

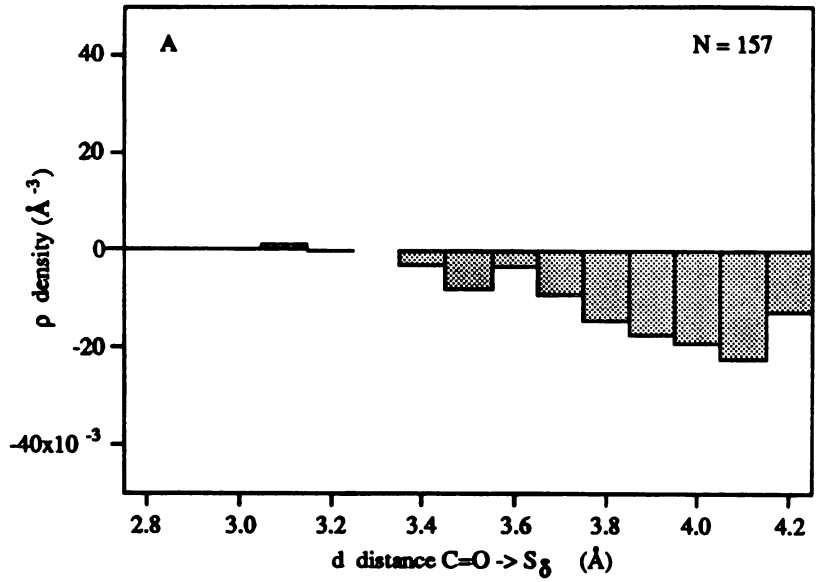
Methionine

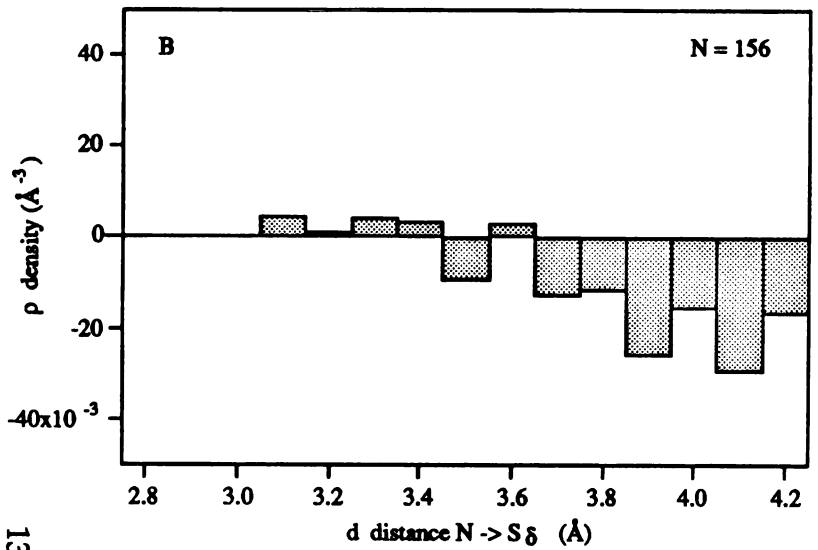
Hydrogen bonding is not particularly prevalent among methionine residues. In fact, as shown by the mostly negative distance difference distribution in Figure 5.2B, there are *fewer* nitrogen atoms near methionine Sδ's, on average, than there are carbon atoms. This situation undoubtedly arises both because methionine is hydrophobic and thus surrounded mostly by carbon atoms, and because Sδ is usually more than 5 Å from the backbone. The distance difference distribution for carbonyl oxygen is very similar to the nitrogen distribution (Fig. 5.2A) even though one would expect to find more short distances to nitrogen since nitrogen can donate hydrogen bonds while carbonyl oxygen is a hydrogen bond acceptor like Sδ of methionine.

Close approaches between hydroxyl groups of serine, threonine, and tyrosine occur with greater frequency than close contacts to carbon, particularly in the expected hydrogen-bonding range of 3.0 to 3.6Å (Fig. 5.2C). Beyond 3.6Å, however, the frequency of carbon interactions is higher. This result suggests that methionine-Sδ -- HO- hydrogen bonds are not negatively biased by x-ray crystallographic refinement schemes since few are found even in the range of van der Waals contact distance where they would be if refinement did not allow for closer contacts. We did not find any angular preferences for methionine-Sδ -- HO- hydrogen bonds.

An upper limit on the frequency with which methionine participates in hydrogen bonding may be estimated by computing the number of methionine Sδ's which are within 4.0 Å of either a hydroxyl group or a nitrogen. Of the methionine Sδ's in our data set, twenty-five percent were within 4.0 Å of a hydroxyl oxygen or nitrogen (see Table V).

A good example of a hydrogen bond to methionine can be found in the combined neutron and x-ray structure of myohemerythrin (2MHR) (Sheriff et al., 1987) between the hydroxyl group of Thr 110 and the sulfur of Met 76 (Figure 5.3). The Sδ - $H\gamma_1$


A
N = 157
40
20
0
-20
-40x10 -3
ρ density (Å -3)
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
d distance C=O -> Sδ (Å)


B
N = 156
40
20
0
-20
-40x10 -3
ρ density (Å -3)
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
d distance N -> Sδ (Å)

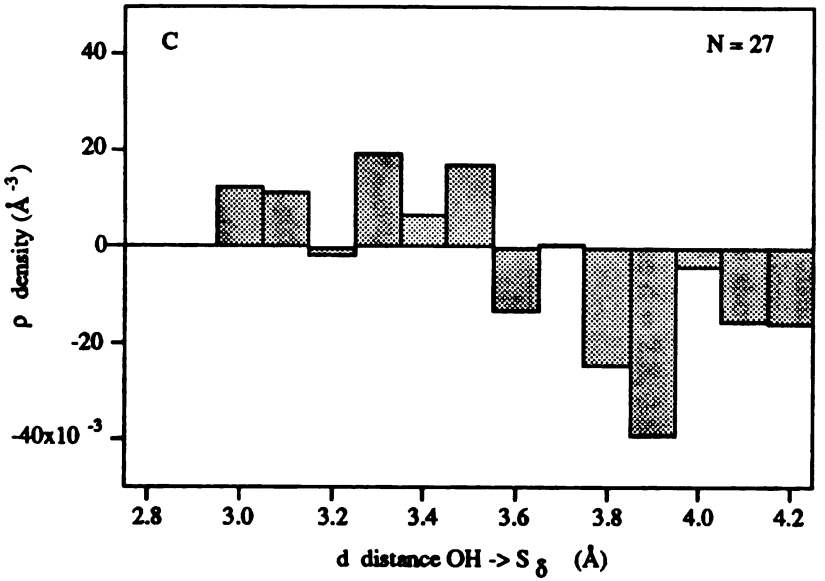

C
N = 27
40
20
0
-20
-40x10 -3
ρ density (Å -3)
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
d distance OH -> Sδ (Å)

Figure 5.2. Distance difference distributions for methionine. A) carbonyl oxygen to methionine-Sδ distances B) nitrogen to methionine-Sδ distances and C) hydroxyl oxygen to methionine-Sδ distances. The number of instances found (N) of each type of interaction is shown in the top right-hand corner. These distributions were prepared in the following manner.

for each d where d = 2.8, 2.9,...,4.2 Å, the donor/acceptor density per residue, ρ(d), is given by:

$$\rho(d) = \left(\frac{n_{X->S}(d)}{N_X} - \frac{n_{C->S}(d)}{N_C} \right) \frac{N_{tot}}{N_{met} \cdot V_{shell}(d)}$$

and $$V_{shell}(d) = \frac{4\pi\left((d + 0.05)^3 - (d - 0.05)^3\right)}{3}$$

where $n_{X->S}(d)$ is the number of atoms or functional groups of type X (e.g. -OH, carbonyl oxygen,...) near sulfur at a distance d
$n_{C->S}(d)$ is the number of carbon atoms near sulfur at a distance d
N_x is the total number of atoms of type X in the data set
N_C is the total number of carbons
N_{tot} is the total number of atoms in the data set
N_{met} is the total number of methionines

The density, ρ(d), therefore, may be interpreted as the excess or deficit in hydrogen bond partners at a particular distance as compared to the carbon atom density. The factor of N_{tot} is an arbitrary scaling factor which scales the donors/acceptors or carbon atoms to be equal to the total number of atoms in the data set.

Table V.

Frequency of vicinity (≤ 4.0 Å) of potential hydrogen bond donor/acceptor groups and carbon

potential donor or acceptor group(s)	Methionine Sδ		Half-Cystine Sγ		Cysteine Sγ	
	n	%	n	%	n	%
-OH	14	5	29	11	13	12
$-NH_n$	71	23	64	24	39	36
>C=0	70	23	158	59	67	62
-OH or $-NH_n$	78	25	85	32	47	43
-OH, $-NH_n$ or >C=0	118	38	177	66	78	72
-C-	241	78	200	75	80	73

Footnote for Table V: This table shows the frequency with which one finds potential hydrogen bond donor or acceptor groups and carbon (**-OH** = hydroxyl; **$-NH_n$** = nitrogen; **>C=O** = carbonyl oxygen; **-C-** = carbon) in the vicinity of the sulfur atoms of methionine, half-cystine, and cysteine. **n** is the number of residues with a sulfur atom within 4.0 Å of at least one member of the donor/acceptor group. **%** is the percentage residues found near the donor/acceptor group. Thus, 29 of the half-cystines in the data set (or 11%) are near at least one hydroxyl group. The data set contains 307 methionines, 268 half-cystines, and 109 cysteines. When more than one group is listed as the donor/acceptor (e.g. "-OH or $-NH_n$") then the number and percentage shown are the number of sulfurs near *either* one donor/acceptor group or the other.

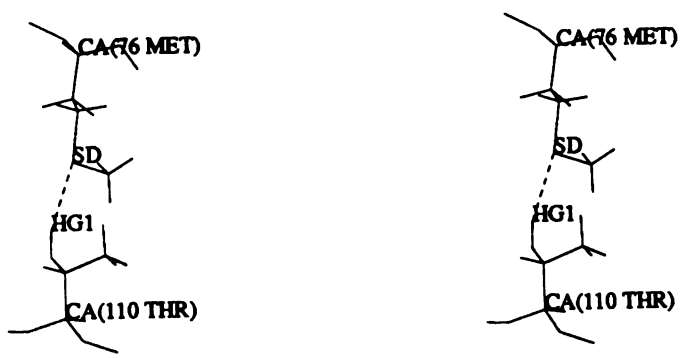


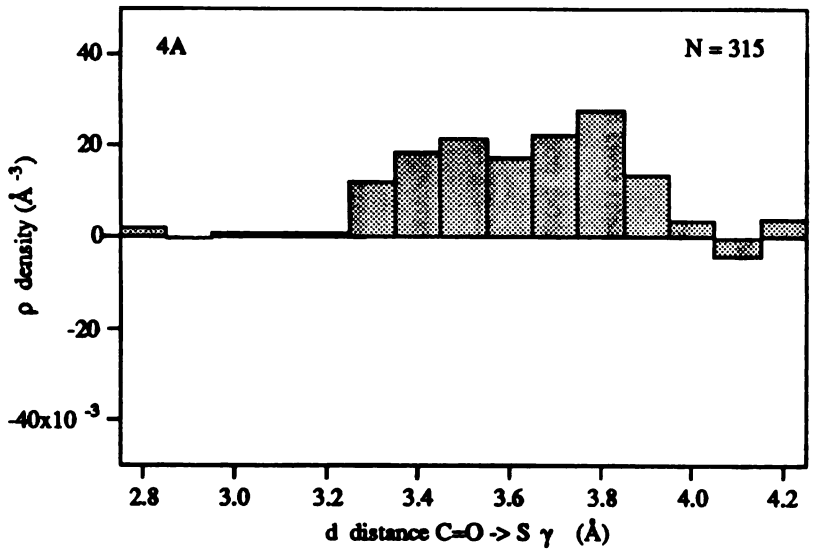
Figure 5.3. Example of a hydrogen bond between the hydroxyl group of Thr 110 and Sδ of Met 76 in myohemerythrin (2MHR).

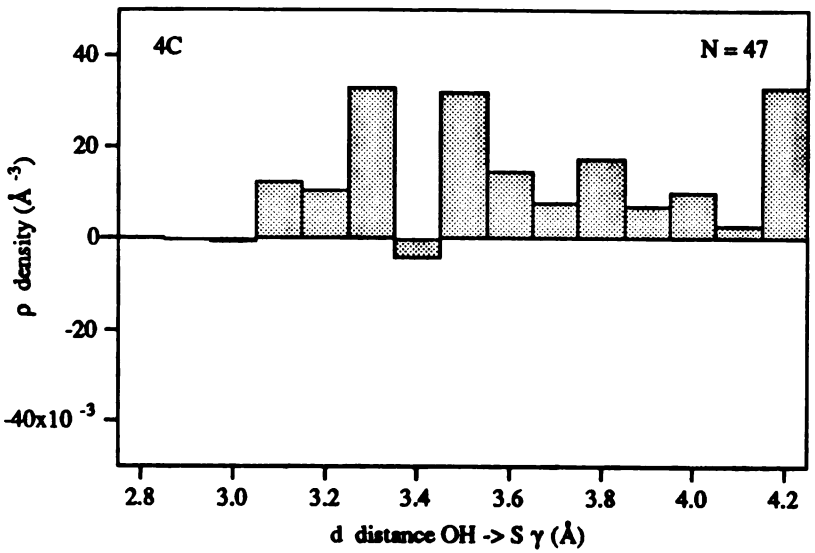
distance is 2.58Å and the Sδ - Oγ1 distance is 3.50Å. The bond is nearly linear with ∠Oγ1 Hγ1 Sδ = 163°. The donor group is not directed at either lone pair of sulfur. Rather, it is in the plane of atoms Cγ, Sδ and Cε closer to Cγ ($\theta \approx 180$; $\phi \approx 50$ for both Oγ1 and Hγ1).

Half-Cystine

As with methionine, short distances between hydroxyl groups and γ-sulfurs of half-cystine residues are observed, suggesting that half-cystine can act as a hydrogen bond acceptor of hydroxyl (Figure 5.4C). Short distances to nitrogen are rarer (Figure 5.4B). Thirty-two percent of half-cystines in our data set had their Sγ within 4.0 Å of a hydroxyl oxygen or nitrogen atom (see Table V). This sets an upper limit for how frequently half-cystine participates in hydrogen bonding in proteins.

Curiously, although hydrogen bonds can not exist between carbonyl oxygen and half-cystine sulfur, we found a significant number of short distances between these groups, in the 3.3 to 4.0 Å range, with a peak at 3.8 Å (Figure 5.4A). Sixteen percent (49 of 302) of carbonyl-O -- Sγ short distances were between the *ith* half-cystine sulfur and the backbone carbonyl oxygen of either the *i - 2* or the *i - 3* residue. Upon examining these interactions further using computer graphics (Ferrin et al., 1988; Jarvis et al., 1988), we noted that many of them occur when the *i - 2* or *i - 3* residue at the end of an α-helix or a β-strand and the *ith* half-cystine residue is in a turn or loop. This result is not surprising if one considers that half-cystine is most commonly observed in coil-type secondary structure and that the polypeptide chain must undergo a 180° chain reversal in order for the topological requirement of the disulfide bond to be satisfied (Thornton, 1981). Apart from the general observation regarding secondary structure, we could not find any other conformational similarity among the examples encountered. There were no angular preferences among the half-cystine Sγ -- donor/acceptor pairs.





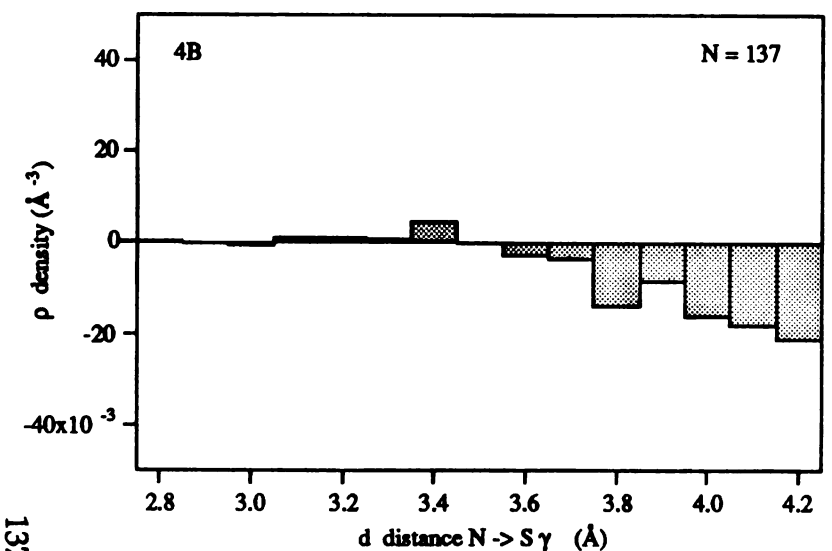
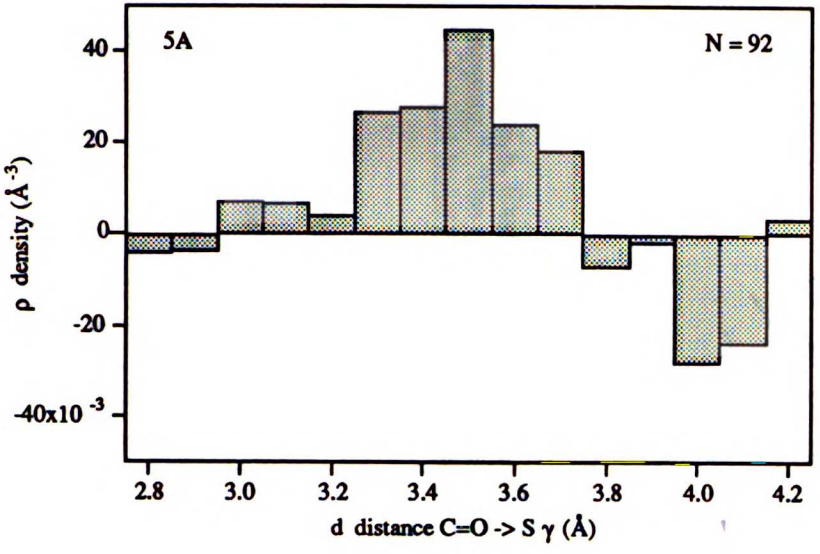


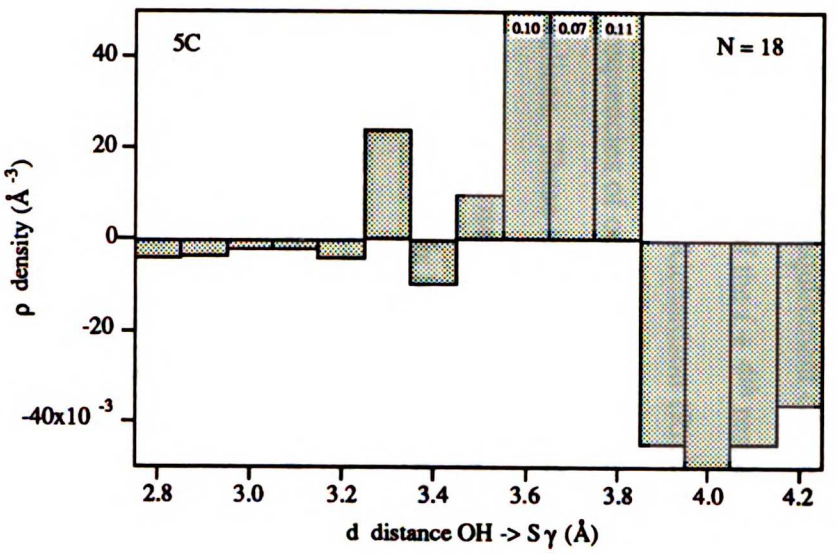
Figure 5.4. Distance difference distributions for half-cystine. Histograms were prepared as described in figure 5.2. **A.** carbonyl oxygen to half-cystine-Sγ; **B.** nitrogen to half-cystine-Sγ; **C.** hydroxyl oxygen to half-cystine-Sγ. The number of instances found (N) of each type of interaction is shown in the top right-hand corner.

Cysteine

Of the three residue types surveyed here, cysteine participates in hydrogen bonding most frequently. Hydrogen bonds between the sulfhydryl group of cysteine and carbonyl oxygen are particularly numerous (Figure 5.5A). Short distances between -SH and nitrogen ($-NH_n$) are common as well and several hydrogen bonds to between -SH and -OH were also observed (Figs. 5.5B, 5.5C). In all, 72 percent of cysteines in our data set of protein structures were found to be within less than 4.0Å of a carbonyl oxygen, nitrogen or hydroxyl oxygen (see Table V). This gives an estimate of the frequency with which cysteine could participate in hydrogen bond formation. Most cysteines (62%) were found near carbonyl oxygens. Proportionately more cysteine sulfurs were found in the vicinity of nitrogen than half-cystine or methionine sulfurs (36% vs. 23% and 24%) suggesting that cystine Sγ also has a greater propensity to behave as a hydrogen bond acceptor. This is also demonstrated in the positive distance distribution for nitrogen (Figure 5.5B). We did not find any angular preferences for cystine-Sγ -- donor/acceptor pairs.

Twenty-seven contacts were found between sulfhydryl of cysteine residue *i* and the carbonyl oxygen of residue *i - 4*. Like serine and threonine (Gray & Matthews, 1984), if the two residues are in a helical conformation, the sulfhydryl group of cysteine can hydrogen bond to the carbonyl oxygen of the *i - 4* residue if cysteine adopts a c_1 angle of -60 degrees. The *i - 4* th carbonyl oxygen is still able to bond to the i^{th} nitrogen and helical geometry does not appear to be compromised. It has been observed that cysteine preferentially adopts a helical conformation -- 47% of cysteines are found in helices (Thornton, 1981). In glycogen phosphorylase, (Newgard et al., 1989; Sprang et al., 1988) 5 of 8 cysteines are in helical conformations and all of these exhibit the *i* --> *i - 4* hydrogen bonding described here. Cysteine residues just beyond the C-termini of helices can also "cap" terminal helical residues three or four residues prior in sequence by forming hydrogen bonds to their carbonyl oxygens. In this manner, the





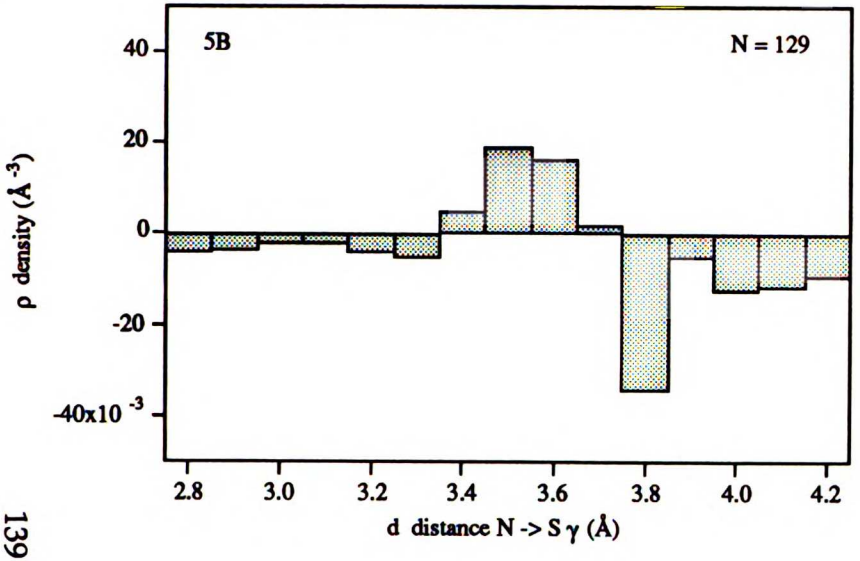


Figure 5.5. Distance difference distributions for cysteine. Histograms were prepared as described in figure 5.2. **A.** carbonyl oxygen to cysteine-Sγ; **B.** nitrogen to cysteine-Sγ; **C.** hydroxyl oxygen to cysteine-Sγ. The number of instances found (N) of each type of interaction is shown in the top right-hand corner.

hydrogen bond requirement of one of the terminal residues is fulfilled even though the amide nitrogen of the $i + 4^{th}$ residue is not available for hydrogen bond formation. Richardson and Richardson (Richardson & Richardson, 1988) have observed greater frequencies of occurence of serine, threonine, and glutamine near the C-termini of helices. These residues are capable of forming side chain - main chain hydrogen bonds. Presta and Rose (Presta & Rose, 1988) have postulated that capping residues are important for helix boundary formation during folding. An example of helix capping may be found in carp parvalbumin (1CPV) (Moews & Kretsinger, 1975) between the sulfhydryl of residue number 18 and the carbonyl oxygen of leucine 15 (Figure 5.6).

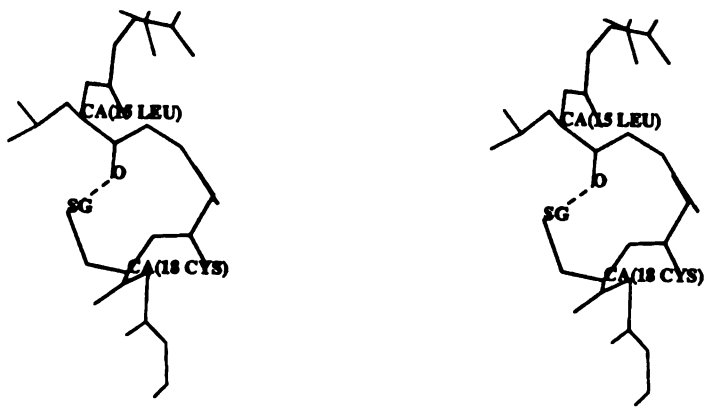


Figure 5.6. Example from carp parvalbumin (1CPV; Moews & Kretsinger, 1975)of "helix capping." The unsatisfied carbonyl oxygen of Leu 15 at the C-terminal of a helix accepts a hydrogen bond from the sulfhydryl group of Cys 18.†

Conclusions

Of the three sulfur-containing amino acids, cysteine participates in hydrogen bonding most frequently. Cysteine is found in the vicinity of hydrogen bond donating

† The structure 1CPV in the Protein Data Bank has been updated. The current structure is 5CPV. This structure is different from 1CPV in that the helix containing residues 15 and 18 is shorter. Residue 18 is not longer in the helix and the putative hydrogen bond distance is 3.84 Å.

or accepting groups 72 percent of the time, most often near carbonyl oxygens. Intrahelical hydrogen bonds between the sulfhydryl group of cysteine and the carbonyl oxygen of the $i - 4^{th}$ residue are quite common, as is C-terminal capping, where the sulfhydryl group of cysteine donates a hydrogen to an unsatisfied carbonyl near the end of the helix. Half-cystine is also frequently found near carbonyl oxygens, though the prevalence of this interaction must be fortuitous and may have more to do with disulfide bond geometry and half-cystine's preference for coil conformation (Thornton, 1981), since half-cystine cannot hydrogen bond to another hydrogen bond acceptor. It is possible that hydrogen bonding ability may influence such factors as side chain conformation and secondary structural preference: intrahelical hydrogen bonding may contribute to cysteine's preference for helical conformation. Surveys of amino acid behavior in proteins should treat free cysteine and half-cystine as unique amino acids.

Sulfur behaves as a hydrogen bond acceptor less frequently. Occasional hydrogen bonds between hydroxyl groups and sulfur are observed in all three amino acid types surveyed (methionine, half-cystine, cysteine.) Hydrogen bonds between the sulfhydryl of cysteine and nitrogen are occasionally observed.

With regard to crystallographic refinement schemes, there does not appear to be a significant bias against short distances between sulfur and potential hydrogen bond donating or accepting groups. If this was the case, we would have observed a cluster of hydrogen bonds at a distance greater than the ideal hydrogen bonding distance. The peak in the distance difference distribution for cysteine-SH -- O-carbonyl is where one would expect it to be -- at 3.5 Å as expected.

While hydrogen bonds to sulfur are not a common feature in globular proteins, their existence should be noted in protein structure modeling schemes and site-directed mutagenesis experiments.

References for Chapter 5

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. (1987). In *Crystallographic Databases -- Information Content, Software Systems, Scientific Applications*. (Allen, F.H., Bergeroff, G. & Sievers, R., ed.), pp. 107-132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Baker, E.N. & Hubbard, R.E. (1984). Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Molec. Biol.* **44**, 97-179.

Bazan, J.F. & Fletterick, R.J. (1988). Viral Cysteine Proteases are Homologous to the Trypsin-like Family of Serine Proteases: Structural and Functional Implications. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 7872-7876.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, 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.

Chakrabati, P. (1989). Geometry of Interaction of Metal Ions with Sulfur-Containing Ligands in Protein Structures. *Biochemistry*. **28**, 6081-6085.

Ferrin, T.E., Huang, C.C., Jarvis, L.E. & Langridge, R. (1988). The MIDAS Display System. *J. Mol. Graphics*. **6**, 13-37.

Gray, T.M. & Matthews, B.W. (1984). Intrahelical Hydrogen Bonding of Serine, Threonine, and Cysteine Residues Within α-Helices and its Relevance to Membrane-bound Proteins. *J. Mol. Biol.* **175**, 75-81.

Jarvis, L., Huang, C., Ferrin, T. & Langridge, R. (1988). UCSF MIDAS: Molecular Interactive Display And Simulation. *J. Mol. Graphics*. **6**, 2-27.

Kamphuis, I.G., Drenth, J. & Baker, E.N. (1985). Thiol Proteases. Comparative Studies Based On The High-Resolution Structures Of Papain And Actinidin, And

On Amino Acid Sequence Information for Cathepsins B And H, And Stem Bromelain. *J. Mol. Biol.* **182**, 317-329.

Kerr, K.A., Ashmore, J.P. & Koetzle, T.F. (1975). A Neutron Diffraction Study of L-Cysteine. *Acta Cryst. Sect. B.* **31**, 2022-2026.

Kollman, P., McKelvey, J., Johansson, A. & Rothenberg, S. (1975). Theoretical Studies of Hydrogen-Bonded Dimers. Complexes involving HF, H_2O, NH_3, HCl, H_2S, Ph_3, HCN, HNC,HCP, CH_2NH, H_2CS,H_2CO, CH_4, CF_3H, C_2H_2, C_2H_4, C_6H_6, F^-, AND H_3O^+. *J. Am. Chem. Soc.* **97**, 955-965.

McGrath, M.E., Wilke, M.E., Higaki, J.N., Craik, C.S. & Fletterick, R.J. (1989). Crystal Structures of Two Engineered Thiol Trypsins. *Biochemistry.* **28**, 9264-9270.

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**, 201-225.

Newgard, C.B., Hwang, P.K. & Fletterick, R.J. (1989). The family of glycogen phosphorylases: structure and function. *Crit. Rev. Biochem. Mol. Biol.* **24**, 69-99.

Presta, L.G. & Rose, G.D. (1988). Helix Signals in Proteins. *Science.* **240**, 1632-1641.

Richardson, J.S. & Richardson, D.C. (1988). Amino acid Preferences for specific Locations at the Ends of α Helices. *Science.* **240**, 1648-1652.

Sheriff, S., Hendrickson, W.A. & Smith, J.L. (1987). Structure of Myohemerythrin in the Azidomet State at 1.7/1.3 Angstroms Resolution. *J. Mol. Biol.* **197**, 273-296.

Sprang, S.R., Acharya, K.R., Goldsmith, E.J., Stuart, D.I., Varvill, K., Fletterick, R.J., Madsen, N.B. & Johnson, L.N. (1988). Structural changes in glycogen phosphorylase induced by phosphorylation. *Nature.* **336**, 215-221.

Thornton, J.M. (1981). Disulphide Bridges in Globular Proteins. *J. Mol. Biol.* **151**, 261-287.

Wilkinson, A.J., Fersht, A.R., Blow, D.M. & Winter, G. (1983). Site-Directed Mutagenesis as Probe of Enzyme Structure and Catalysis: Tyrosyl-tRNA Synthetase Cysteine-35 to Glycine-35 Mutation. *Biochemistry*. **22**, 3581-3586.

Chapter 6.

Unusual Packing of Proteases

Introduction

Proteins are observed to be well-packed (Richards, 1977). One measure of packing density which has recently been used with reference to cubic lattice models of compact polymers is the number of topological contacts (Chan & Dill, 1990) -- the total number intrachain contacts between residues which are not connected in sequence. While investigating the packing density in proteins using this measure, I noted that proteases contain an unusually large number of close contacts between alpha-carbons when compared to other proteins. They also appear to have slightly lower surface area : molecular weight ratios. Here I present these results and propose why proteases exhibit these unusual packing characteristics.

Methods

A data set of 72 protein structures from the Brookhaven Protein Data Bank (PDB) (Bernstein,1977 ; Abola, 1987) was constructed previously (Gregoret & Cohen, 1990). This set includes 18 proteases, among them three bacterial trypsin-like serine proteases (2ALP, 2SGA, 3SGB), six mammalian trypsin-like serine proteases (1TON, 2PKA, 3EST, 3RP2, 4CHA, 4PTP), three aspartyl proteases (2APP, 3APR, 4APE), subtilisin and proteinase K (1SBT, 2PRK), papain and actinidin (9PAP, 2ACT), thermolysin (3TLN), and carboxypeptidase A (5CPA). These cover a variety of structural and mechanistic classes, and a size range of 18 to 34 kilodaltons.

For each protein in the set, the number of intrachain contacts was determined. A contact is defined if the C^{α} - C^{α} distance between residues i and j (for $j > i+2$) is less than or equal to 5.5 Å. Surface areas were calculated using the method of Lee and Richards (Lee & Richards, 1971) as implemented by Handschumacher and Richards in the program ACCESS. Voronoi volumes (Richards, 1974) were calculated using the program VOLUME, also by Handschumacher and Richards. The principle moments of inertia of the data set proteins were also computed.

Results/Discussion

Amongst a data set of 72 protein structures, the18 proteases are amongst the top 28 proteins in terms of contacts per residue (Figure 6.1). The most compact non-proteases in the set of 54 have no obvious structural similarity to the proteases. These are superoxide dismutase (2SOD), cytochrome C' (2CCY), *L7/L12* small ribosomal protein (1CTF), gamma crystallin (1GCR), and hemoglobin (2HHB).

A calculation of the solvent accessible surface areas of the data set proteins gives a similar result (Figure 2): proteases are on the fringe of the expected accessible surface area for their sizes. This aberration may have been missed in previous studies of surface area dependence on protein size because these studies focused on smaller sets of proteins or were limited to monomeric proteins (Chothia, 1975; Janin, 1979; Bryant *et al.*, 1990). It is notable that in these papers, proteases frequently dominated the high molecular weight members of the data sets analyzed. Some of the debate over the exact power law dependence of surface area to molecular weight may owe to the peculiarities of protease packing.

The proteases are not unusually dense: the amino acid packing densities are normal when computed using the method of constructing Voronoi polyhedra. This argues that the forces operating to stabilize the folded protein, such as van der Waals attractions and hydrogen bonds are not present in excess.

One way in which proteases may achieve a smaller surface area : molecular weight ratio simply by being more spherical. The "roundness" of a protein may be computed by

$$\text{Roundness} = \frac{\sqrt{I_{xx} + I_{yy}}}{\sqrt{2 \cdot I_{zz}}}$$

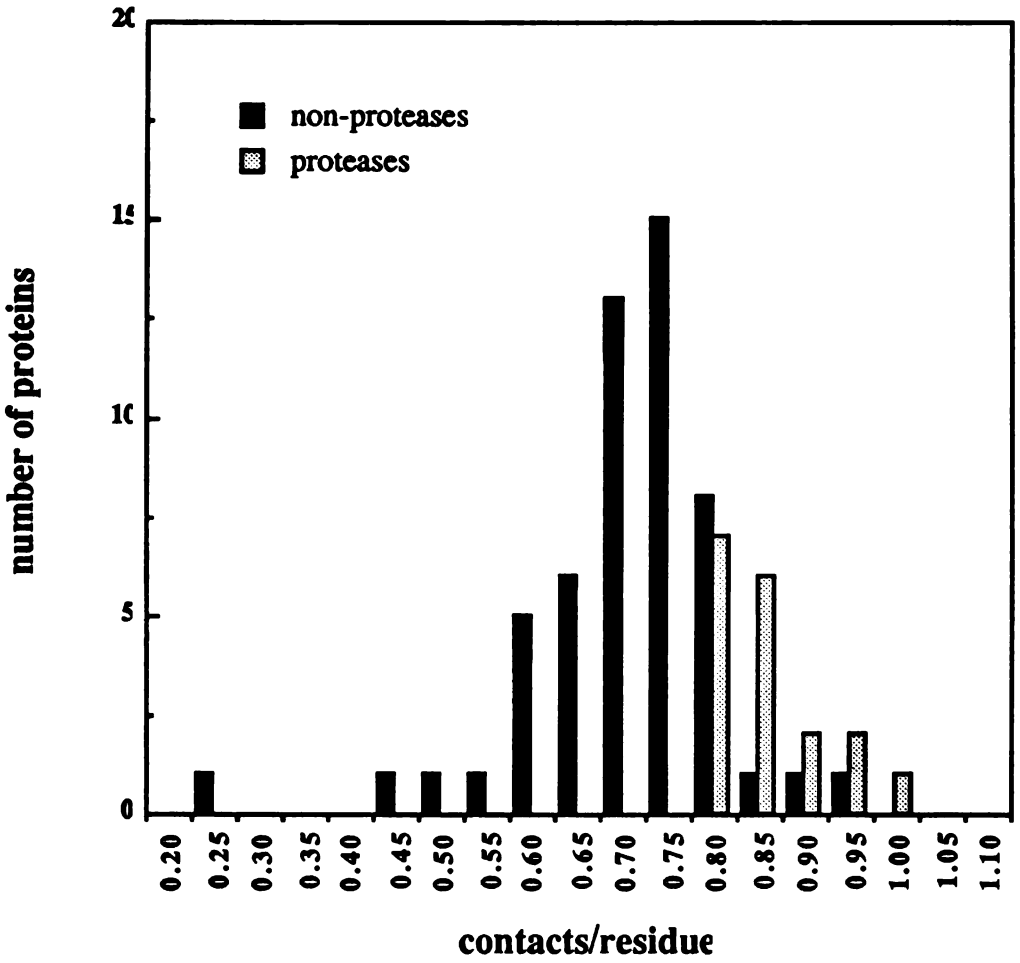
where I_{xx}, I_{yy}, and I_{zz} are the eigenvalues of the inertia tensor

In the limiting case of a perfectly spherical protein, this ratio is equal to 1.0.

Although we normally think of proteases as being oblong, two-domain entities with a binding cleft, several of the structural classes of proteases studied are actually quite round (Figure 6.3). Subtilisin is, in fact, the second most spherical protein in the set, with a roundness of 0.92. Proteinease K is also quite spherical, as is carboxypeptidase A and the bacterial and mammalian serine proteases. The aspartyl proteases, papain, actinidin, and thermolysin are more oblong. These proteins may acheive a high number of intrachain contacts and lower surface areas by other mechanisms, perhaps by having shorter or more well-packed loops on the surface.

Conclusion

I speculate that the large group of proteases may have convergently evolved to be more compact as protection against autolysis and cleavage by other proteases in the same milieu. There appear to be at least two mechanisms by which proteases minimize their surface areas while maximizing the number of intrachain contacts. One way in which proteases acheive this is by becoming more spherical. Not all proteases, however, are round. Some may achieve lower surface area : molecular weight ratios by making their surfaces less convoluted. This could be achieved through shorter or more well-packed loops. This and other alternatives have yet to be investigated in greater detail.



6.1. Distribution of contacts/residue for the data set of 72 proteins. The average number of non-local (greater than three residues apart in sequence) C^{α} - C^{α} contacts for the 54 non-proteases in the data set is 0.75 ± 0.12 contacts/residue. For the 18 proteases, the average is 0.86 ± 0.06 contacts/residue.

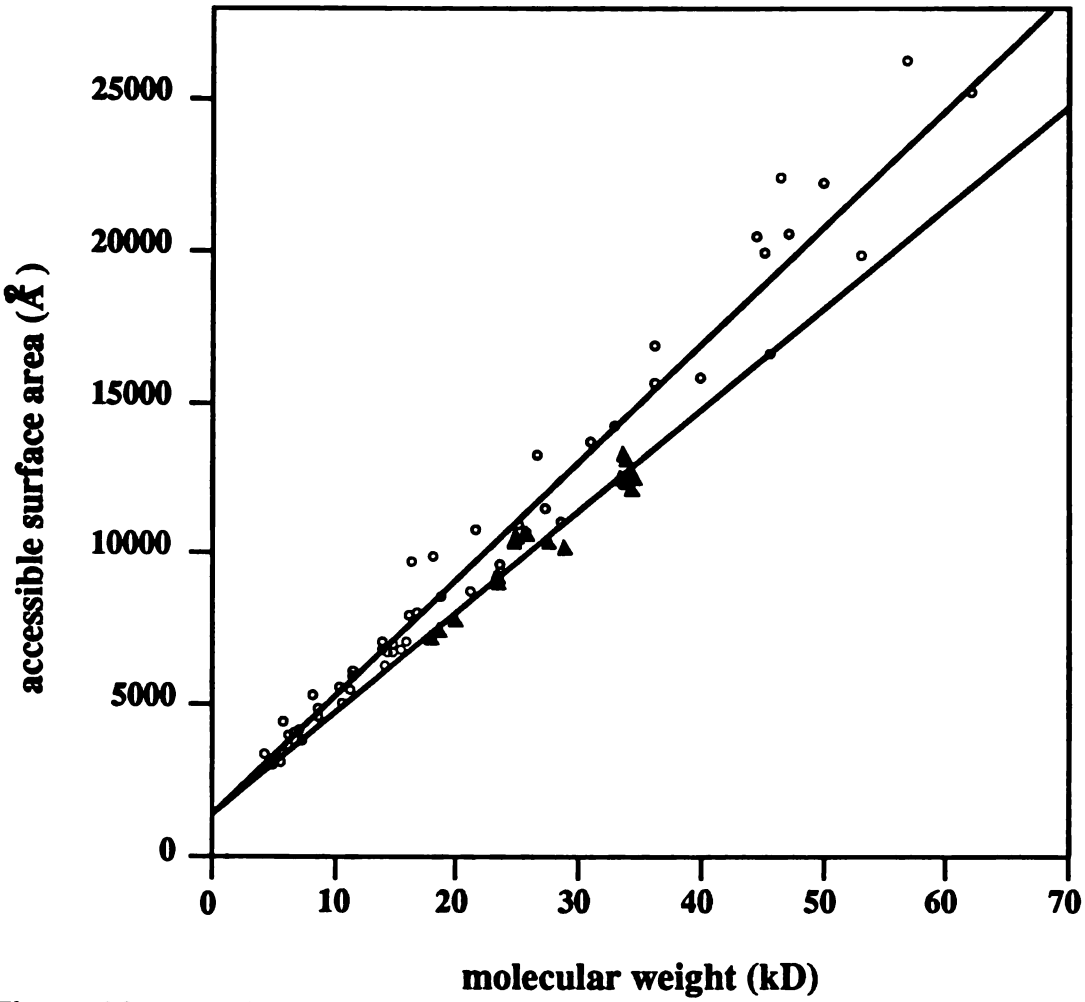


Figure 6.2. Accessible surface area plotted as a function of molecular weight. Surface area was calculated using the method of Lee and Richards. Proteases are designated as filled triangles and other proteins as open circles. Linear equations have been fitted to each set of proteins. The correlation coefficients are r = 0.96 (proteases) and r = 0.99 (non-proteases).

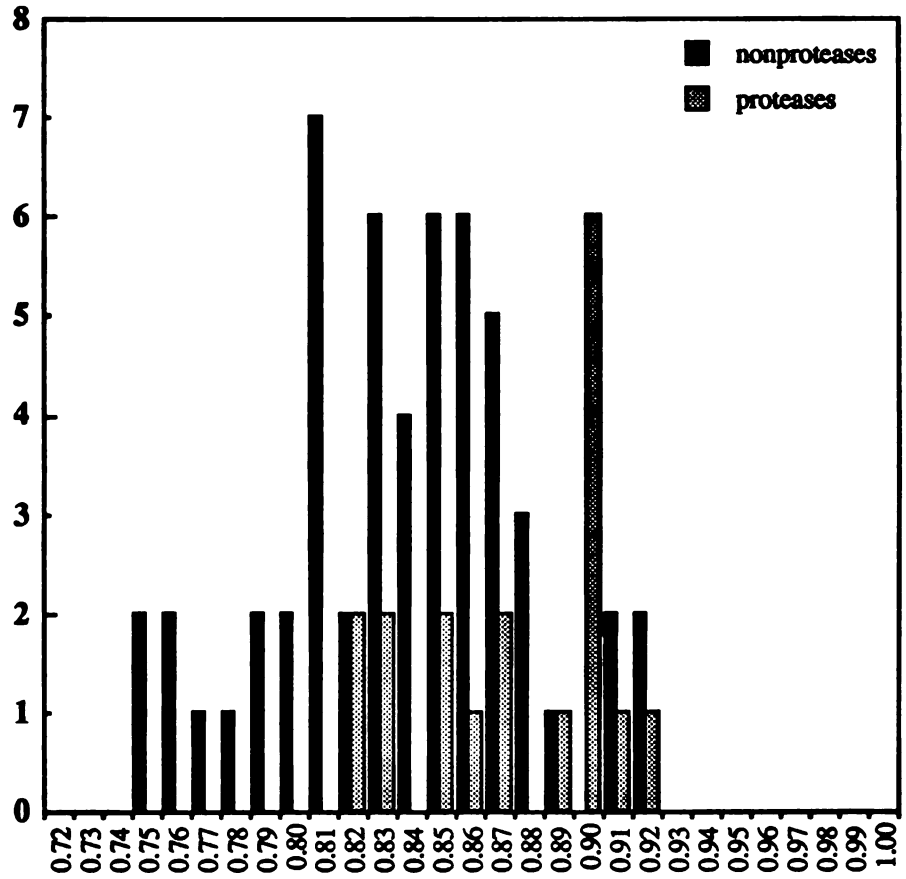


Figure 6.3 Roundness of nonproteases and proteases.

References for Chapter 6

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. & Weng, J. (1987). In *Crystallographic Databases -- Information Content, Software Systems, Scientific Applications*. (Allen, F.H., Bergeroff, G. & Sievers, R., ed.), pp. 107-132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Jr., E.F.M., 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.

Bryant, S.H., Islam, S.A. & Weaver, D.L. (1989). The Surface Area of Monomeric Proteins: Significance of Power Law Behavior. *Proteins: Struct., Func., Genet.* **6**, 418-423.

Chothia, C. (1975). Structural Invariants in Protein Folding. *Nature (London).* **254**, 304-308.

Chan, H.S. & Dill, K.A. (1990). Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA.* **87**, 6388-6392.

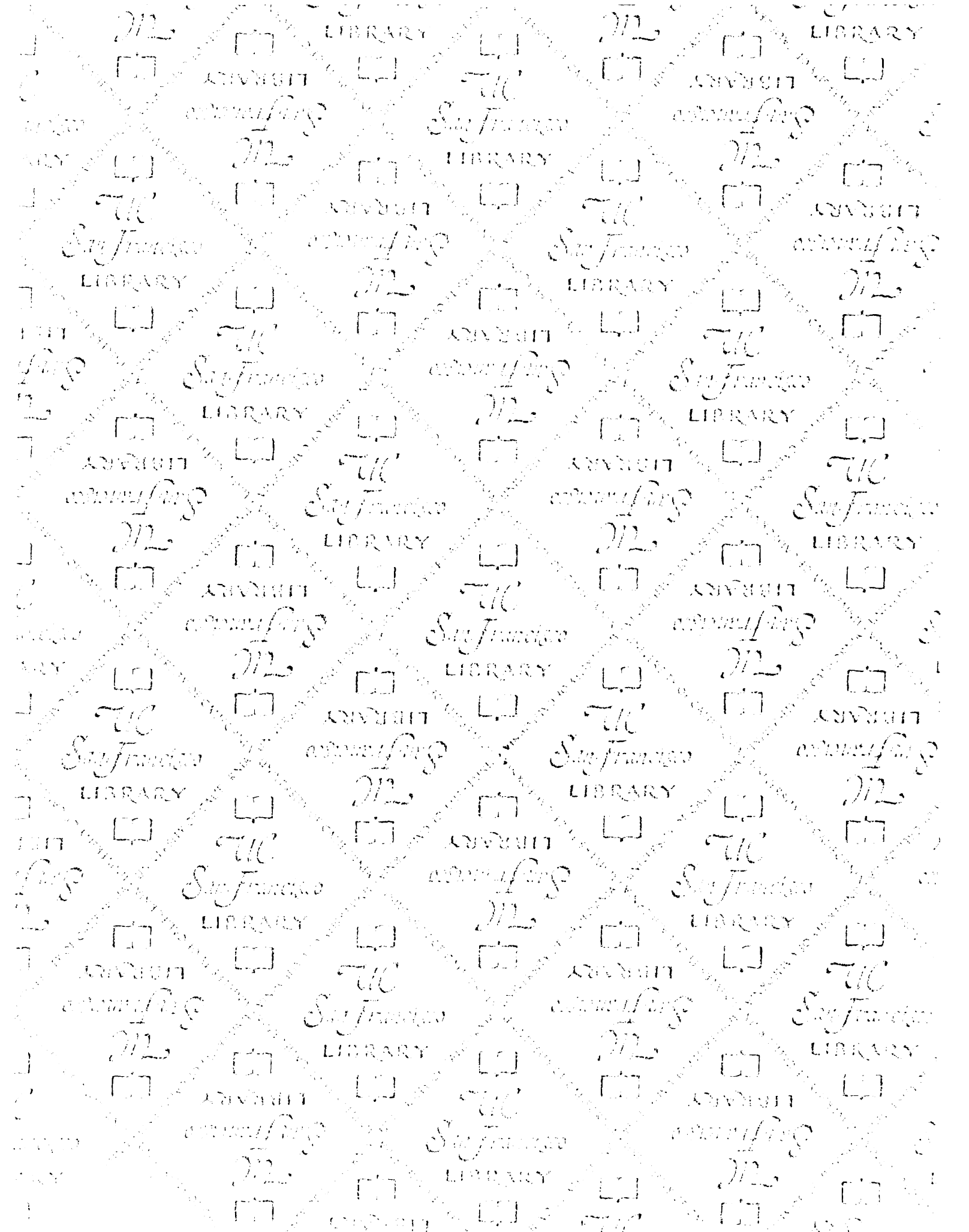
Gregoret, L.M. & Cohen, F.E. (1990). Novel Method for the Rapid Evaluation of Packing in Protein Structures. *J. Mol. Biol.* **211**, 959-974.

Janin, J. (1979) Surface and Inside Volumes in Globular Proteins. *Nature.* **277**, 491-492.

Lee, B. & Richards, F.M. (1971). The Interpretation of Proteins Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **55**, 379-400.

Richards, F.M. (1974). The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density. *J. Mol. Biol.* **82**, 1-14.

Richards, F.M. (1977). Areas, Volumes, Packing and Protein Structure. *Ann. Rev. Biophys. Bioeng.* **6**, 151-176.


LIBRARY